Инженерный вестник Дона, №5 (225)
ivdon.ru/ru/magazine/archive/n5y2025/10038

Комплексный анализ русскоязычных текстов на основе нейросетевых моделей трансформерного типа

В.И. Шиян[1], В.Н. Марков[2]

[1]Кубанский государственный университет, Краснодар
[2]Кубанский государственный технологический университет, Краснодар

Аннотация: Статья посвящена комплексному анализу русскоязычных текстов с использованием нейросетевых моделей, в основу которых положен двунаправленный кодировщик представлений трансформера (Bidirectional Encoder Representations from Transformers – BERT). В работе применяются специализированные модели для русского языка: RuBERT-tiny, RuBERT-tiny2 и RuBERT-base-cased. Предложенный метод охватывает морфологический, синтаксический и семантический уровни анализа, включая лемматизацию, определение частей речи, морфологических признаков, синтаксических отношений, семантических ролей и связей. Использование моделей семейства BERT позволяет достичь точности выше 98% для лемматизации, 97% для определения частей речи и морфологических признаков, 96% для синтаксического анализа и 94% для семантического анализа. Метод подходит для задач, требующих глубокого понимания текста, и может быть оптимизирован для работы с большими корпусами.
Ключевые слова: русскоязычные тексты, морфологический анализ, синтаксический анализ, семантический анализ, лемматизация, RuBERT, обработка естественного языка.

Введение

Обработка естественного языка (Natural Language Processing – NLP) является одной из ключевых задач в области искусственного интеллекта. Основная цель такой обработки заключается в преобразовании текста в структурированную информацию, которая может быть использована для анализа, классификации и интерпретации текста [1]. В данной статье рассматривается задача комплексного анализа русскоязычных текстов, включающая морфологический, синтаксический и семантический уровни. Проблемными факторами являются развитая морфология, словоизменение и богатая грамматика русского языка. Область применения включает реферирование, определение семантической близости текстов, определение первоисточника, кластеризацию текстов и другие задачи, связанные с обработкой текстовой информации.

Существующие методы обработки текста часто ограничиваются отдельными аспектами анализа, такими как морфологический или синтаксический разбор, что не позволяет глубоко понять смысл текста. Традиционные подходы к лемматизации и определению частей речи основаны на правилах, не всегда учитывающих контекст. Современные методы, такие как двунаправленный кодировщик представлений трансформера (Bidirectional Encoder Representations from Transformers – BERT), предлагают более гибкие решения [2], но их применение для русского языка требует адаптации. В целом, существующие решения часто имеют недостаточную точность в семантическом анализе и высокую ресурсоёмкость, особенно для больших текстовых корпусов.

Многоуровневый метод анализа текстов на основе BERT

Предлагаемый метод анализа текстов на основе BERT включает последовательное использование специализированных модулей (рис. 1). На первом этапе слова преобразуются в числовые векторные представления – эмбеддинги [3]. Для этого можно использовать предварительно обученные модели, такие как RuBERT-tiny, RuBERT-tiny2 или RuBERT-base-cased [4], которые учитывают слово и его окружение, обеспечивая высокую точность.

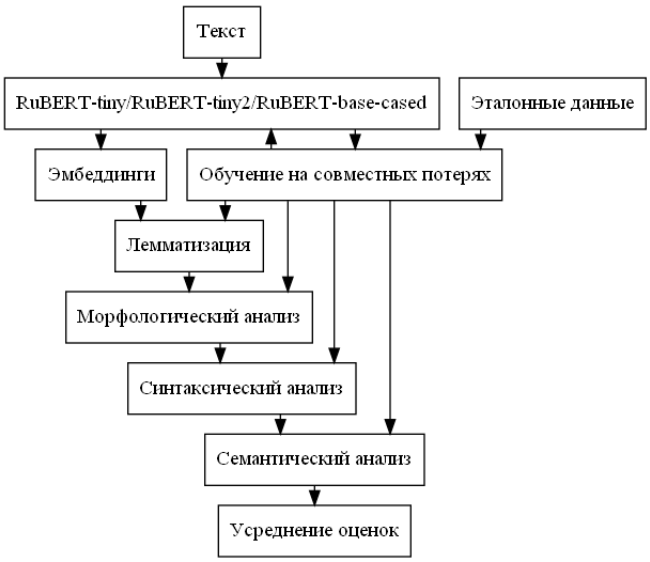

Рис. 1. – Схема многоуровневого анализа текста

Следующим этапом является лемматизация, на котором определяется нормальная форма слова. Затем выполняется морфологический анализ, в рамках которого определяется часть речи и морфологические характеристики каждого слова. После этого проводится синтаксический анализ с использованием метода биаффинной классификации [5], который выявляет синтаксические отношения между словами. На заключительном этапе выполняется семантический анализ, где определяются семантические роли и связи между словами. BERT используется на этапах: преобразование в эмбеддинги, лемматизация, морфологический анализ, синтаксический анализ и семантический анализ.

Все модули, за исключением усреднения оценок, объединены общим механизмом обучения, основанным на совместных потерях. Совместные потери представляют собой подход, при котором модель оптимизирует несколько целевых функций одновременно, учитывая вклад каждой задачи. Это позволяет модели обучаться сразу на всех задачах, что повышает точность и согласованность анализа.

Для валидации эффективности метода на каждом уровне анализа разработаны метрики, учитывающие лингвистические особенности русского языка. Они позволяют количественно оценить точность обработки и согласованность результатов между уровнями.

Оценка точности лемматизации

Функция ScoreLemma используется для экспертной оценки точности лемматизации, формула (1):

$$\begin{aligned}&\text{ScoreLemma}\left(test,\ gold\right)=\\&=\text{LemmaWeight}\left(gold_{POS}\right)\cdot\left[\text{Normalize}\left(test_{lemma}\right)=\text{Normalize}\left(gold_{lemma}\right)\right],\end{aligned}\quad(1)$$

где LemmaWeight – вес леммы, зависящий от части речи POS в эталонных данных $gold$; Normalize – функция нормализации, которая приводит строку

к нижнему регистру и заменяет символ ё на е; $[x = y]$ – скобки Айверсона, принимающие значение 1, если условие истинно, и 0, если оно ложно.

Формула учитывает вес леммы в зависимости от части речи [6]: вес неизменяемых слов, таких, как предлоги или союзы, ниже, чем у изменяемых слов, например, существительных или прилагательных. Если леммы совпадают после нормализации, которая включает приведение к нижнему регистру и замену символов, то оценка умножается на вес леммы, иначе оценка равна нулю.

Вес леммы LemmaWeight определяется экспертно и учитывает часть речи слова, формула (2):

$$\text{LemmaWeight}(word) = \\ = \begin{cases} \alpha, \text{ если } word \text{ – предл., союз, межд., част., зн. препин., символ,} \\ \beta, \text{ иначе,} \end{cases} \quad (2)$$

где $\alpha = 0,3$ – вес служебных частей речи и символов; $\beta = 0,7$ – вес остальных частей речи.

Оценка точности определения части речи и морфологических признаков

Функция ScorePOS оценивает точность определения частей речи. Она сравнивает части речи в тестовых данных с эталонными, формула (3):

$$\text{ScorePOS}(test,\ gold) = [test_{POS} = gold_{POS}], \quad (3)$$

где $test_{POS}$ – часть речи в тестовых данных; $gold_{POS}$ – часть речи в эталонных данных.

Функция ScoreFeats, используемая для оценки морфологических признаков, определяется формулой (4):

$$\text{ScoreFeats}\left(test,\ gold\right)=$$
$$\text{Penalty}\left(test_{feats},\ gold_{feats}\right)\cdot\frac{\sum\limits_{Cat\in gold_{feats}}\text{FeatsWeight}\left(Cat\right)\cdot\left[test_{feats}^{Cat}=gold_{feats}^{Cat}\right]}{\sum\limits_{Cat\in gold_{feats}}\text{FeatsWeight}\left(Cat\right)},\quad(4)$$

где Penalty – штраф за избыточность морфологических признаков в тестовых данных; FeatsWeight – вес морфологической категории, который равен количеству уникальных значений в данной категории.

Эта функция показывает, насколько точно определены морфологические признаки слова в тестовых данных по сравнению с эталонными. Если все морфологические признаки совпадают, оценка равна 1. Если морфологические признаки не совпадают, оценка равна 0.

Штраф за избыточность морфологических признаков в тестовых данных определяется с помощью функции Penalty, которая учитывает количество морфологических категорий в тестовых и эталонных данных, формула (5):

$$\text{Penalty}\left(x,\ y\right)=\begin{cases}\dfrac{1}{1+\text{Length}\left(x\right)-\text{Length}\left(y\right)},\ \text{если}\ \text{Length}\left(x\right)>\text{Length}\left(y\right),\\ 1,\ \text{иначе},\end{cases}\quad(5)$$

где $\text{Length}\left(x\right)$ – количество морфологических категорий в тестовых данных; $\text{Length}\left(y\right)$ – количество морфологических категорий в эталонных данных.

Вес морфологической категории определяется с помощью функции FeatWeight, формула (6):

$$\text{FeatsWeight}\left(Category\right)=\left|Category\right|.\quad(6)$$

Эта функция учитывает количество уникальных значений в данной категории. Например, если категория падеж имеет 8 значений, то её вес будет равен 8.

Оценка синтаксических отношений

Для оценки синтаксических отношений используется функция оценки немаркированных связей (Unlabeled Attachment Score – UAS), которая показывает, насколько правильно определены синтаксические отношения в предложении без учёта их типа, формула (7):

$$\mathrm{UAS}(test,\ gold) = [test_{head} = gold_{head}], \tag{7}$$

где $test_{head}$ – индекс слова, от которого зависит текущее слово, в тестовых данных; $gold_{head}$ – индекс слова, от которого зависит текущее слово, в эталонных данных.

Если индекс слова, от которого зависит текущее слово в тестовых данных совпадает с эталонным, оценка равна 1. Если индексы не совпадают, оценка равна 0.

Для оценки синтаксических отношений с учётом их типа используется функция оценки маркированных связей (Labeled Attachment Score – LAS), которая показывает, насколько правильно определены синтаксические отношения в предложении с учётом их типа, формула (8):

$$\mathrm{LAS}(test,\ gold) = [test_{head} = gold_{head}] \cdot [test_{deprel} = gold_{deprel}], \tag{8}$$

где $test_{head}$ – индекс слова, от которого зависит текущее слово, в тестовых данных; $gold_{head}$ – индекс слова, от которого зависит текущее слово, в эталонных данных; $test_{deprel}$ – тип синтаксического отношения в тестовых данных; $gold_{deprel}$ – тип синтаксического отношения в эталонных данных.

Если индекс слова, от которого зависит текущее слово, и тип синтаксического отношения в тестовых данных совпадают с эталонными, оценка равна 1. Если хотя бы один из параметров не совпадает, оценка равна 0.

Оценка семантических ролей и связей

Для оценки семантических ролей используется функция ScoreSemrole, которая показывает, насколько корректно определены семантические роли слов в предложении в тестовых данных по сравнению с эталонными, формула (9):

$$\text{ScoreSemrole}\left(test,\ gold\right)=\left[test_{semrole}=gold_{semrole}\right], \tag{9}$$

где $test_{semrole}$ – семантическая роль в тестовых данных; $gold_{semrole}$ – семантическая роль в эталонных данных.

Если семантические роли совпадают, оценка равна 1, иначе – 0.

Для оценки семантических связей используется функция ScoreSemrel [7], которая определяет, насколько корректно установлены семантические связи между словами в тестовых данных по сравнению с эталонными, формула (10):

$$\text{ScoreSemrel}\left(test,\ gold\right)=\frac{1}{1+\text{Distance}\left(test_{semrel},\ gold_{semrel}\right)}, \tag{10}$$

где Distance – расстояние между узлами, которые представляют семантические связи, в дереве.

Если семантические связи совпадают, оценка равна 1. Если связи находятся на одной ветви, оценка уменьшается в зависимости от расстояния между ними. Если связи находятся на разных ветвях дерева, оценка равна 0.

Указанное расстояние определяется с помощью функции Distance, которая вычисляет длину пути между узлами, формула (11):

$$\text{Distance}\left(u,\ v\right)=\begin{cases}\text{PathLength}\left(u,\ v\right), & \text{если } u \text{ и } v \text{ в одном дереве,}\\ \infty, & \text{если } u \text{ и } v \text{ в разных деревьях,}\end{cases} \tag{11}$$

где PathLength – длина пути между узлами u и v в дереве через ближайшего общего предка, формула (12):

$$\text{PathLength}\left(u,\ v\right)=\text{Depth}\left(u\right)+\text{Depth}\left(v\right)-2\cdot\text{Depth}\left(\text{LCA}\left(u,\ v\right)\right), \tag{12}$$

где $\mathrm{LCA}(u, v)$ – ближайший общий предок узлов u и v.

Усреднение оценок

Для усреднения оценок ScoreX лемм, частей речи, морфологических признаков, синтаксических отношений, синтаксических отношений с типами, семантических ролей или семантических связей используется функция AverageScoreX, формула (13):

$$\mathrm{AverageScoreX}(tests, golds) = \frac{\sum_{(test, gold) \in (tests, golds)} \mathrm{ScoreX}(test, gold)}{\sum_{gold \in golds} \mathrm{ScoreX}(gold, gold)}, \quad (13)$$

где ScoreX – одна из оценок, рассчитанных функциями оценки; *tests* – тестовые данные; *golds* – эталонные данные.

Описание эксперимента

Для обучения и тестирования модели использовался корпус текстов на русском языке, состоящий из 10000 предложений. Общий объём корпуса составляет около 1 миллиона слов. Корпус содержит трёхуровневую разметку: морфологическую, включающую леммы, части речи и морфологические признаки; синтаксическую, охватывающую синтаксические отношения между словами; и семантическую, включающую семантические роли и связи [8, 9].

Разметка данных выполнена в формате CoNLL-U, который является стандартом для лингвистических исследований и обработки текстов. В корпусе представлены 17 частей речи, 15 морфологических признаков, 41 тип синтаксических отношений, 133 семантические связи и 860 семантических ролей. Формат CoNLL-U позволяет хранить подробную информацию о каждом слове в предложении, включая его лемму, часть речи, морфологические признаки, синтаксические отношения, а также семантические роли и связи.

Результаты обработки текстов представлены в формате CoNLL-U, который является стандартом для лингвистических данных и хранит информацию о каждом слове в предложении, включая его лемму, часть речи, морфологические признаки, синтаксические отношения, семантические роли и связи.

Для оценки эффективности предложенного метода был проведён эксперимент с использованием трёх моделей: RuBERT-tiny, RuBERT-tiny2 и RuBERT-base-cased. Результаты эксперимента показали (рис. 2), что модель RuBERT-base-cased демонстрирует наилучшие результаты во всех задачах [10], достигая точности выше 98% для лемматизации, 97% – для определения частей речи и морфологических признаков, 96% – для синтаксического анализа и 94% – для семантического анализа.

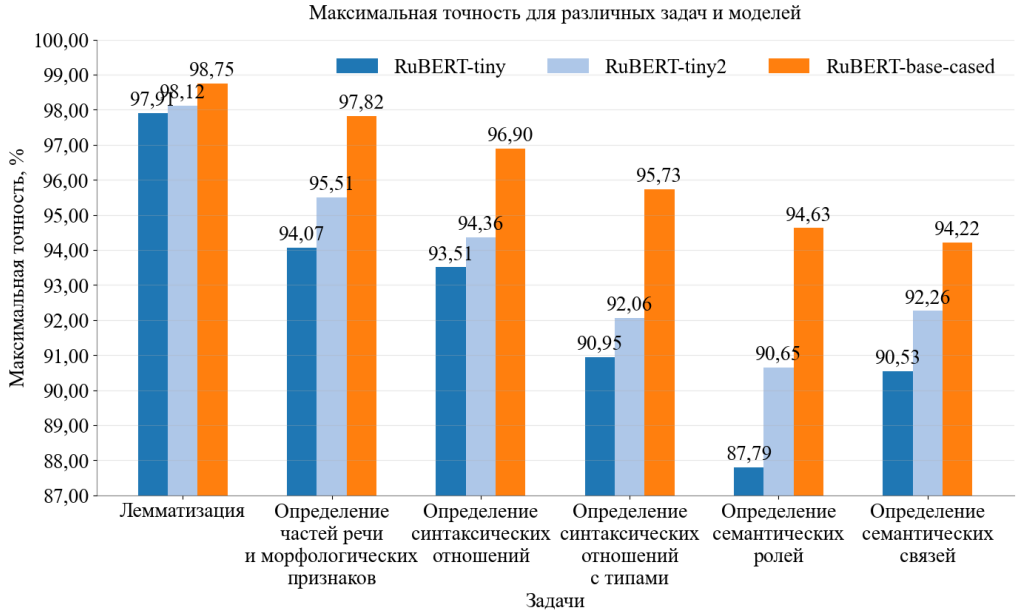

Рис. 2. – Максимальная точность для различных задач и моделей

Модель RuBERT-base-cased лидирует по всем показателям, что связано с её более сложной архитектурой и большим количеством параметров. Однако модели RuBERT-tiny и RuBERT-tiny2 также показывают хорошие результаты, особенно в задачах, где требуется менее глубокий анализ контекста.

Для наглядности результаты обработки текста представлены в виде сети фреймов (рис. 3), построенной на данных формата CoNLL-U, для предложения «Редкая птица долетит до середины Днепра!». Сеть отображает морфологические, синтаксические и семантические характеристики слов, включая лемму, часть речи, морфологические признаки, синтаксические отношения, семантические роли и связи.

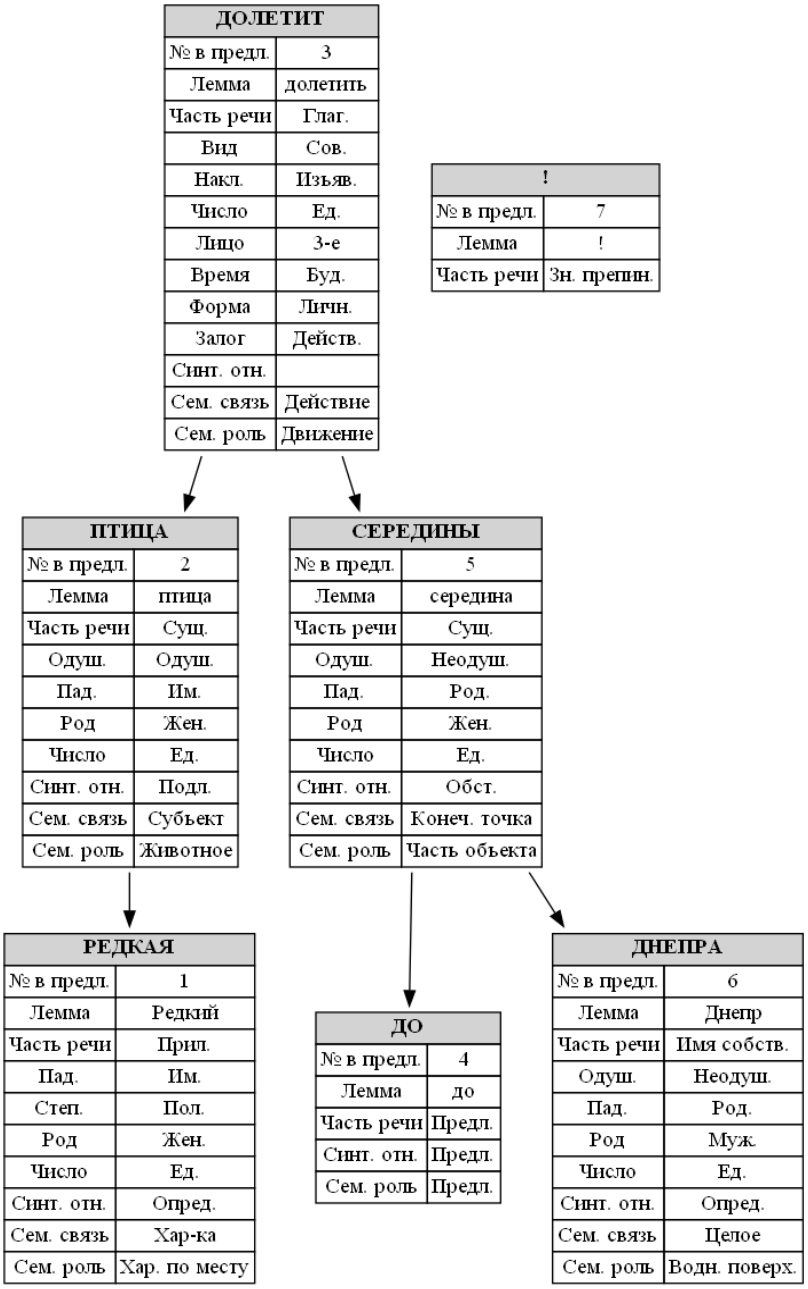

Рис. 3. – Сеть фреймов для предложения «Редкая птица долетит до середины Днепра!»

В этом предложении каждое слово анализируется на морфологическом уровне: определяются его лемма, часть речи и грамматические признаки. Например, слово «Редкая» приводится к лемме «Редкий», классифицируется как прилагательное и описывается признаками: именительный падеж, положительная степень, женский род, единственное число. Синтаксический анализ фразы «птица долетит» выявляет отношения между словами: «птица» становится подлежащим, а «долетит» – корневым глаголом, от которого зависят остальные элементы предложения. На семантическом уровне словам присваиваются роли и связи: «птица» определяется как субъект действия, а «середины» – как конечная точка. В результате текст раскладывается на структурированные компоненты, что позволяет глубже понять его смысл.

Стоит отметить, что нейросетевые модели, несмотря на высокую точность, например, 98% для лемматизации, могут допускать ошибки в редких или неоднозначных случаях. Примером может служить ошибочное преобразование слова «долетит» в «долетить». Такие ошибки связаны с вероятностной природой моделей и ограничениями обучающих данных, которые могут не охватывать все возможные контексты и формы слов.

Заключение

Предложенный метод выделяется своей комплексностью, охватывая все уровни анализа текста: морфологический, синтаксический и семантический. Использование моделей на основе BERT, таких как RuBERT-tiny, RuBERT-tiny2 и RuBERT-base-cased, позволяет эффективно решать задачи на каждом из этих уровней. В частности, модель RuBERT-base-cased демонстрирует наилучшие результаты во всех задачах, достигая точности выше 98% для лемматизации, 97% для определения частей речи и морфологических признаков, 96% для синтаксического анализа и 94% для семантического анализа.

Комплексный подход, объединяющий несколько уровней анализа в единую систему, позволяет не только повысить точность на каждом этапе, но и обеспечить согласованность результатов между различными уровнями. Это особенно важно для задач, требующих глубокого понимания текста, таких как определение семантических ролей и связей между словами. Модели RuBERT-tiny и RuBERT-tiny2 также показывают хорошие результаты, особенно в задачах, где требуется менее глубокий анализ контекста, что делает их подходящими для приложений, где важны скорость обработки и экономия ресурсов.

Таким образом, предложенный метод представляет собой эффективное решение для комплексного анализа русскоязычных текстов, сочетающее в себе высокую точность и возможность глубокого понимания структуры и смысла текста. В дальнейшем планируется исследовать возможности оптимизации модели для работы с большими текстовыми корпусами, а также улучшение точности в задачах семантического анализа.

Литература

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. URL: arxiv.org/abs/1810.04805.

2. Rogers A., Kovaleva O., Rumshisky A. A Primer in BERTology: What We Know About How BERT Works // Transactions of the Association for Computational Linguistics. 2020. Vol. 8. Pp. 842-866.

3. Лыченко Н.М., Сороковая А.В. Сравнение эффективности методов векторного представления слов для определения тональности текстов // Математические структуры и моделирование. 2019. №4. С. 97-110.

4. Николенко С. Transformer: внимание на себя. СПб.: НИУ ВШЭ, 2019. 77 с.

5. Гринчук А.В. Матричные и тензорные разложения в задачах обработки естественного языка: дис. ... канд. физ.-мат. наук: 05.13.18. М., 2021. 98 с.

6. Сапин А.С. Построение нейросетевых моделей морфологического и морфемного анализа текста // Труды ИСП РАН. 2021. Т. 33. №4. С. 117-130.

7. Егунова А.И., Комаров Р.С., Федосин С.А., Шибайкин С.Д., Жарков Р.А. Сравнительный анализ методов извлечения знаний из текстов для построения онтологий // Инженерный вестник Дона. 2025. №2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9823.

8. Шишаев М.Г. Нейросетевые модели в задачах семантического анализа текстов на естественном языке // Труды Кольского научного центра РАН. 2020. №8-11. URL: cyberleninka.ru/article/n/neyrosetevye-modeli-v-zadachah-semanticheskogo-analiza-tekstov-na-estestvennom-yazyke.

9. Аксенов К.А., Сунь Л. Эволюция и современное состояние систем ответов на вопросы: технологии распознавания намерений и именованных сущностей с использованием модели BERT // Инженерный вестник Дона. 2024. №7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9371.

10. Салып Б.Ю., Смирнов А.А. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка // StudNet. 2022. №5. URL: cyberleninka.ru/article/n/analiz-modeli-bert-kak-instrumenta-opredeleniya-mery-smyslovoy-blizosti-predlozheniy-estestvennogo-yazyka.

References

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. URL: arxiv.org/abs/1810.04805.

2. Rogers A., Kovaleva O., Rumshisky A. A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics. 2020. Vol. 8. Pp. 842-866.

3. Lychenko N.M., Sorokovaya A.V. Matematicheskie struktury i modelirovanie. 2019. №4. p. 97-110.

4. Nikolenko S. Transformer: vnimanie na sebya [Transformer: Self-Attention]. SPb.: NIU VShE, 2019. 77 p.

5. Grinchuk A.V. Matrichnye i tenzornye razlozheniya v zadachah obrabotki estestvennogo jazyka [Matrix and Tensor Decompositions in Natural Language Processing Tasks]: dis. ... kand. fiz.-mat. nauk: 05.13.18. M., 2021. 98 p.

6. Sapin A.S. Trudy ISP RAN. 2021. T. 33. №4. pp. 117-130.

7. Egunova A.I., Komarov R.S., Fedosin S.A., Shibajkin S.D., Zharkov R.A. Inzhenernyj vestnik Dona. 2025. №2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9823.

8. Shishaev M.G. Trudy Kol'skogo nauchnogo centra RAN. 2020. №8-11. URL: cyberleninka.ru/article/n/neyrosetevye-modeli-v-zadachah-semanticheskogo-analiza-tekstov-na-estestvennom-yazyke.

9. Aksenov K.A., Sun' L. Inzhenernyj vestnik Dona. 2024. №7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9371.

10. Salyp B.YU., Smirnov A.A. StudNet. 2022. №5. URL: cyberleninka.ru/article/n/analiz-modeli-bert-kak-instrumenta-opredeleniya-mery-smyslovoy-blizosti-predlozheniy-estestvennogo-yazyka.

Дата поступления: 18.02.2025

Дата публикации: 25.04.2025