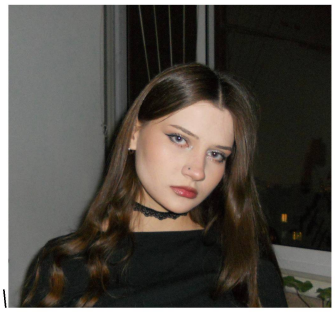
УДК 004.912:004.8

TRANSFORMER-АРХИТЕКТУРА: КАК ИЗМЕНИЛА ПОДХОД К РАЗРАБОТКЕ СИСТЕМ ВЫЯВЛЕНИЯ ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ

К.С. Крез
Ассистент кафедры проектирования информационно-компьютерных систем БГУИР,
магистр технических наук
k.krez@bsuir.by

М.А. Кривоносова
Студентка 3 курса БГУИР Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)
krvria@gmail.com



А.Р. Шипуль
Студентка 3 курса БГУИР Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)
angelinashipul@gmail.com

А.С. Гугалев
Студент 3 курса БГУИР Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)
gugalevandrei@gmail.com

К.С. Крез
Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: корреляционный анализ цифровых следов пользователей, нейронные сети, разработка и анализ структуры хранения данных.

М.А. Кривоносова
Студентка 3 курса Белорусского государственного университета информатики и радиоэлектроники Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)

А.Р. Шипуль
Студентка 3 курса Белорусского государственного университета информатики и радиоэлектроники Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)

А.С. Гугалев

Студент 3 курса Белорусского государственного университета информатики и радиоэлектроники Факультета компьютерного проектирования по специальности Информационные системы и технологии (в бизнес-менеджменте)

Аннотация. В статье рассматривается влияние архитектуры Transformer на развитие методов автоматического выявления текстовых заимствований в современных системах антиплагиата. Показано, что переход от лексико-статистических и строковых методов сравнения к контекстно-зависимому семантическому анализу способствует повышению качества обнаружения перефразированных и частично модифицированных заимствований. Проанализированы ограничения традиционных подходов, основанных на шинглах, n-граммах и статических эмбеддингах, а также раскрыта роль механизма self-attention в формировании контекстуальных представлений текстовых фрагментов. Особое внимание уделено применению архитектурных стратегий bi-encoder и cross-encoder в составе гибридного конвейера антиплагиатной проверки, обеспечивающего баланс между производительностью и точностью. Представлен упрощённый пример семантического сопоставления сегментов документа с локальным хранилищем и расчёта коэффициента семантических совпадений. Сделан вывод о целесообразности интеграции трансформерных моделей в многоуровневые системы антиплагиата, сочетающие точные лексические алгоритмы и семантический анализ.

Ключевые слова: семантическое сходство текстов, STS, Transformer, self-attention, BERT, SBERT, sentence embeddings, семантический поиск.

Введение. В условиях стремительного роста объёмов цифровых текстов задача автоматического выявления заимствований перестала сводиться к поиску буквальных совпадений. Современная система антиплагиата должна обнаруживать не только дословное копирование, но и перефразирование, замену слов синонимами, перестановку фрагментов, а также частичное изменение формулировок при сохранении исходного смысла [1]. Именно поэтому при разработке таких систем всё большую роль играет семантическое сравнение текстов Semantic Textual Similarity (далее – STS), позволяющее оценивать близость фрагментов на уровне смысла, а не только на уровне совпадения слов.

Существовавшие ранее системы обнаружения заимствований на протяжении длительного времени основывались преимущественно на лексических методах, шинглов анализе и строковых алгоритмах сопоставления текстов. Эти подходы остаются полезными для выявления прямых заимствований, однако их эффективность заметно снижается в случаях смыслового переписывания текста. Появление архитектуры Transformer стало важным этапом развития интеллектуальных систем проверки текстовых работ, поскольку позволило перейти от поверхностного сопоставления к контекстно-зависимому анализу содержания. В результате современные антиплагиатные системы приобретают способность учитывать семантическую близость текстовых фрагментов, что существенно расширяет их функциональные возможности в образовательной и научной сфере.

Основная часть. Развитие систем автоматической проверки текстов на заимствования приоритет отдавался методам, ориентированным на поиск точных или частично изменённых совпадений текстовых последовательностей. Для этого применялись алгоритмы строкового сравнения, n-граммный анализ, шингли и лексико-статистические методы, в том числе Term Frequency – Inverse Document Frequency (далее – TF-IDF). Такие решения достаточно эффективно выявляли прямое копирование и заимствования с минимальными изменениями. Однако при перефразировании, перестановке слов или замене отдельных лексических единиц их точность заметно снижалась. По этой причине тексты, выражающие одно и то же содержание разными языковыми средствами, часто оценивались системой как различные.

Использование векторных представлений слов, таких как Word2Vec, GloVe и FastText [2], стало значимым шагом в развитии методов анализа текстового сходства. Эти модели позволили лучше учитывать семантическую близость слов и терминов, благодаря чему повысилась эффективность обработки текстов, связанных общей тематикой. Однако их ключевым ограничением оставался статический характер представлений: каждому слову соответствовал один и тот же вектор независимо от контекста. В задачах выявления заимствований это создавало риск неточной интерпретации многозначных слов, особенно в

научно-технических текстах, где смысл языковой единицы определяется не только самим словом, но и его окружением.

Дальнейшее развитие методов обработки текста сопровождалось применением рекуррентных нейронных сетей и моделей Long Short-Term Memory (далее – LSTM), позволивших в большей степени учитывать порядок слов и структуру последовательности. Однако высокая вычислительная сложность и ограниченная эффективность при обработке длинных текстов затрудняли их использование в высоконагруженных антиплагиатных системах, ориентированных на анализ больших массивов документов.

В системах автоматической проверки текстов стали появлятся архитектуры Transformer, в основе которой лежит механизм self-attention. Его особенность состоит в том, что при обработке текста каждый токен рассматривается не изолированно, а с учётом связи со всеми остальными элементами последовательности [3]. Для задач обнаружения заимствований это особенно важно, так как смысл фрагмента нередко сохраняется даже после изменения структуры предложения, замены отдельных слов и частичного перефразирования. Благодаря этому self-attention позволяет глубже учитывать контекст, точнее выявлять смысловые соответствия между фрагментами и более надёжно отличать тематическое сходство от фактического смыслового заимствования.

Появление моделей семейства Bidirectional Encoder Representations from Transformers (далее – BERT) и их производных сделало возможным использование контекстуальных эмбеддингов текстовых фрагментов в системах обнаружения заимствований.

В результате задача антиплагиата постепенно трансформируется из задачи поиска текстовых совпадений в задачу многоуровневого анализа, включающего лексическое, структурное и семантическое сравнение. Трансофрмер-модели выступают в этом процессе как ядро семантического уровня, позволяя обнаруживать скрытые и перефразированные заимствования.

При разработке системы антиплагиата на базе трансформеров ключевым становится выбор архитектурной стратегии сравнения фрагментов текста. В практических системах применяются два базовых подхода – bi-encoder и cross-encoder, каждый из которых решает свою часть задачи [4].

В схеме bi-encoder каждый сегмент проверяемого документа и каждый сегмент фрагмента источников кодируются независимо в фиксированный вектор. Далее сравнение выполняется в векторном пространстве, как правило, с использованием косинусной меры. Такой подход особенно эффективен на этапе предварительного поиска подозрительных совпадений, поскольку эмбеддинги фрагмента можно вычислить заранее и сохранить в индексе. Это позволяет оперативно находить семантически близкие фрагменты даже при работе с крупными локальными хранилищами текстовых работ.

В архитектуре cross-encoder сегмент проверяемого документа и найденный текст-кандидат подаются в модель как единая пара. Такой подход даёт возможность механизму внимания напрямую учитывать связи между двумя фрагментами и за счёт этого точнее оценивать степень их сходства. Хотя данный подход требует больших вычислительных затрат по сравнению с bi-encoder, он особенно полезен на этапе уточняющей проверки, где необходимо более точно оценить ограниченное число отобранных кандидатов и уменьшить количество ложноположительных результатов.

С точки зрения практической реализации в системах обнаружения заимствований наиболее оправданным является применение двухэтапной архитектуры. На первом этапе модель типа bi-encoder обеспечивает быстрый отбор потенциально релевантных кандидатов для каждого сегмента документа. На втором этапе модель типа cross-encoder выполняет уточнённую оценку степени семантической близости между сопоставляемыми фрагментами. Подобная организация процесса позволяет совместить высокую производительность с повышенной точностью анализа, что имеет принципиальное значение при массовой проверке студенческих и научных работ.

Практическая часть. В рамках разработки антиплагиатной системы семантическое сравнение целесообразно рассматривать как один из этапов общего конвейера анализа текстового документа. На начальной стадии выполняется предварительная обработка, включающая извлечение текстового содержимого, удаление служебных элементов, нормализацию и сегментацию документа на смысловые фрагменты. Далее каждый выделенный сегмент преобразуется в векторное представление с использованием модели типа bi-encoder, после чего осуществляется поиск семантически близких фрагментов в локальном хранилище или в базе ранее проверенных работ [5].

Отобранные кандидаты могут передаваться в модуль уточняющего сравнения, реализованный, например, на основе архитектуры cross-encoder, где степень релевантности оценивается с более высокой точностью. На основании полученных оценок формируется карта совпадений по сегментам проверяемого документа. На следующем этапе модуль может вычислять интегральный показатель потенциальных заимствований с учётом количества подтверждённых совпадений, длины соответствующих фрагментов и характера их распределения по структуре документа. Такой подход обеспечивает возможность не только выявления отдельных потенциально заимствованных участков, но и формирования интерпретируемого итогового отчёта, предназначенного для преподавателя, эксперта или иного уполномоченного пользователя.

Ниже может быть приведён упрощённый пример программной реализации семантического модуля антиплагиатной системы. В рамках данного примера сегменты проверяемого документа и сегменты локального хранилища кодируются в эмбеддинги, после чего для каждого сегмента документа определяется наиболее близкий по смыслу сегмент из хранилища. Если значение меры сходства превышает заданный порог, соответствующее совпадение интерпретируется как потенциально релевантное семантическое совпадение. На этой основе может быть рассчитан упрощённый коэффициент семантических совпадений, определяемый как доля сегментов документа, для которых были обнаружены семантически близкие соответствия.

```
    import torch
    import torch.nn.functional as F
    from transformers import AutoTokenizer, AutoModel
    model_name = "google-bert/bert-base-multilingual-cased"
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    model = AutoModel.from_pretrained(model_name)
    def mean_pooling(model_output, attention_mask):
        token_embeddings = model_output.last_hidden_state
        mask                =                attention_mask.unsqueeze(-
1).expand(token_embeddings.size()).float()
        return  torch.sum(token_embeddings  *  mask,  dim=1)  /
torch.clamp(mask.sum(dim=1), min=1e-9)
    def encode_texts(texts):
        encoded = tokenizer(
            texts,
            padding=True,
            truncation=True,
            return_tensors="pt"
        )
        with torch.no_grad():
            output = model(**encoded)
        embeddings              =              mean_pooling(output,
encoded["attention_mask"])
        embeddings = F.normalize(embeddings, p=2, dim=1)
```

```
        return embeddings
    document_segments = [
        "Transformer использует механизм self-attention для анализа
контекста.",
    "Система должна находить не только дословные, но и
перефразированные совпадения."
    ]
    reference_segments = [
        "Механизм self-attention в Transformer позволяет учитывать
связи между токенами.",
    "Система обязана выявлять перефразирование и скрытые
совпадения."
    ]
    doc_emb = encode_texts(document_segments)
    ref_emb = encode_texts(reference_segments)
    threshold = 0.70
    matched = 0
    for i, text in enumerate(document_segments):
        scores = torch.matmul(doc_emb[i], ref_emb.T)
        best_idx = torch.argmax(scores).item()
        score = scores[best_idx].item()
        if score >= threshold:
            matched += 1
        status = "релевантно" if score >= threshold else "не
подтверждено"
        print(f"Исходный сегмент: {text}")
        print(f"Найденный                                    сегмент:
{reference_segments[best_idx]}")
        print(f"Сходство: {score:.4f}")
        print(f"Статус: {status}")
        print("-" * 80)
    print(f"\nДоля    релевантных    совпадений:    {matched    /
len(document_segments):.2%}")
```

С целью повышения точности анализа и снижения количества ложноположительных результатов в системе обнаружения заимствований целесообразно предусмотреть второй этап проверки, реализованный на основе архитектуры cross-encoder. В отличие от bi-encoder, где сегменты сравниваются через заранее вычисленные векторы, cross-encoder получает пару фрагментов одновременно и оценивает их релевантность с учётом прямого межтекстового взаимодействия токенов. Такой подход вычислительно дороже, поэтому он применяется только к ограниченному числу кандидатов, найденных на первом этапе семантического поиска. В результате система получает более точную фильтрацию подозрительных совпадений и более надёжную оценку степени семантической близости сравниваемых фрагментов.

```
    from sentence_transformers import CrossEncoder
    cross_encoder = CrossEncoder("cross-encoder/ms-marco-MiniLM-L-
6-v2")
    pairs = [[m["segment"], m["matched_segment"]] for m in matches]
    cross_scores = cross_encoder.predict(pairs)
    for m, s in zip(matches, cross_scores):
    m["cross_score"] = float(s)
```

```
print(matches[0]["segment"], "->", matches[0]["cross_score"])
```

После этого итоговое решение о подтверждении совпадения может приниматься не только по порогу bi-encoder, но и с учётом cross_score, что особенно полезно при анализе перефразированных фрагментов, близких по теме, но не являющихся фактическим заимствованием.

Ограничения и особенности применения трансформеров в антиплагиате. Несмотря на высокую эффективность, применение моделей класса Transformer в системах обнаружения заимствований сопряжено с рядом ограничений. Во-первых, наличие семантической близости между фрагментами не всегда свидетельствует о некорректном заимствовании. В научных и учебных текстах закономерно встречаются типовые определения, общеупотребимые формулировки и устойчивые терминологические конструкции, совпадение которых по смыслу не может рассматриваться как достаточное основание для вывода о наличии заимствования. В связи с этим итоговое решение должно приниматься на основе совокупности признаков, а не исключительно на значении семантической меры.

Во-вторых, качество обнаружения в значительной степени определяется предметной областью и характеристиками фрагмента, в рамках которого функционирует система. Модель, демонстрирующая высокую результативность на текстах общего назначения, может менее точно интерпретировать специализированные термины и контексты в технических, юридических или медицинских документах. Это обусловливает актуальность задачи доменной адаптации моделей, а также настройки порогов релевантности с учётом специфики проверяемых текстов.

В-третьих, существенное влияние на результат оказывает стратегия сегментации документа. Использование чрезмерно коротких фрагментов может приводить к потере значимого контекста, тогда как чрезмерно длинные сегменты затрудняют точную локализацию потенциального заимствования. Следовательно, эффективность системы обнаружения заимствований определяется не только выбором нейросетевой модели, но и качеством всей методики обработки текста, включая сегментацию, индексацию, переранжирование кандидатов и расчёт итоговых показателей.

Заключение. Архитектура Transformer оказала существенное влияние на развитие систем обнаружения заимствований, обеспечив переход от выявления буквальных совпадений к обнаружению семантически близких и перефразированных фрагментов, требующих дополнительной экспертной интерпретации. Использование механизма self-attention и контекстуальных эмбеддингов позволило расширить аналитические возможности обработки текста и повысить точность выявления смысловых совпадений в тех случаях, когда традиционные строковые и шингловые методы оказываются недостаточно эффективными. Это имеет особое значение при проверке учебных и научных работ, в которых заимствование нередко маскируется посредством замены лексических единиц, перестановки синтаксических конструкций и частичной переработки исходного текста.

С практической точки зрения наиболее целесообразным решением для реальных систем обнаружения заимствований является применение гибридной архитектуры, в рамках которой лексические методы, bi-encoder и cross-encoder функционируют как взаимодополняющие компоненты единого аналитического конвейера. Подобная организация обеспечивает высокую скорость первичного поиска потенциально подозрительных совпадений, более точную семантическую верификацию отобранных кандидатов и возможность формирования интерпретируемого итогового отчёта по результатам анализа. Вместе с тем следует учитывать, что сама по себе семантическая близость не может рассматриваться как достаточный признак некорректного заимствования. В связи с этим итоговая оценка должна формироваться на основе совокупности признаков, включая характер совпадения, тип текстового фрагмента и контекст его использования.

Список литературы

[1] Wahle, J. P. Identifying Machine-Paraphrased Plagiarism / J. P. Wahle [et al.] // Information for a Better World: Shaping the Global Future: 17th International Conference. – Springer, 2022. – С. 393-413.

[2] Krez, K. S. From words to vectors: text vectorization techniques in natural language processing / K. S. Krez // 2. – Минск: БГУИР, 2025. – С. 60-62.

[3] Rogers, A. A Primer in BERTology: What We Know About How BERT Works / A. Rogers, O. Kovaleva, A. Rumshisky // Transactions of the Association for Computational Linguistics. – 2020. – Т. 8. – С. 842-866.

[4] Reimers, N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks / N. Reimers, I. Gurevych // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. – 2019. – С. 3982-3992.

[5] Karpukhin, V. Dense Passage Retrieval for Open-Domain Question Answering / V. Karpukhin [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – С. 6769-6781.

Авторский вклад

Крез Карина Сергеевна – научное руководство, проверка методологии, проведение литературного обзора, пересмотр и редактирование содержания статьи, утверждение окончательного варианта рукописи для публикации.

Шипуль Ангелина Робертовна, **Кривоносова Мария Анатольевна**, **Гугалев Андрей Сергеевич**, – анализ алгоритмических подходов, участие в написании отдельных разделов рукописи, оформление текста и списка литературы.

TRANSFORMER ARCHITECTURE: HOW IT TRANSFORMED THE APPROACH TO DEVELOPING TEXT PLAGIARISM DETECTION SYSTEMS

K.S. Krez
Assistant of the Department of Information and Computer Systems Design, BSUIR, Master of Technical Sciences

M.A. Krivonosova
3rd year student of the BSUIR Faculty of Computer Engineering, specializing in Information Systems and Technologies (in business management)

A.R. Shipul
3rd year student of the BSUIR Faculty of Computer Engineering, specializing in Information Systems and Technologies (in business management)

A.S. Gugalev
3rd year student of the BSUIR Faculty of Computer Engineering, specializing in Information Systems and Technologies (in business management)

Abstract. This article examines the impact of the Transformer architecture on the development of automated text plagiarism detection methods within modern anti-plagiarism systems. It demonstrates that the transition from lexico-statistical and string-based comparison methods to context-aware semantic analysis contributes to improved detection accuracy for paraphrased and partially modified instances of plagiarism. The limitations of traditional approaches – based on shingles, n-grams, and static embeddings – are analyzed, and the role of the self-attention mechanism in generating contextual representations of text fragments is elucidated. Particular attention is devoted to the application of bi-encoder and cross-encoder architectural strategies within a hybrid anti-plagiarism verification pipeline, designed to strike a balance between performance and accuracy. A simplified example is presented illustrating the semantic matching of document segments against a local repository and the subsequent calculation of a semantic similarity score. The article concludes that integrating Transformer-based models into multi-layered anti-plagiarism systems – which combine precise lexical algorithms with semantic analysis – is a highly advisable strategy.

Keywords: semantic similarity of texts, STS, Transformer, self-attention, BERT, SBERT, sentence embeddings, semantic search.