MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 33, NO. 1, 2026
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Applying Large Language Models to Russian-English Word Alignment

D. A. Morozov[1], A. A. Makhova[2], P. V. Dyachenko[2], A. D. Kozerenko[3]

DOI: 10.18255/1818-1015-2026-1-48-61

[1]Novosibirsk State University, Novosibirsk, Russia
[2]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[3]Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received February 5, 2026
Revised February 22, 2026
Accepted February 24, 2026

This paper investigates the task of automatic word alignment in parallel texts, a fundamental step for training machine translation systems, conducting comparative linguistic studies, and creating linguistic resources. Given the scarcity of annotated data for many language pairs, the applicability of Large Language Models (LLMs) becomes particularly relevant due to their high generalization capabilities and ability to solve tasks without extensive fine-tuning on target datasets. This study presents a comparative analysis of the effectiveness of modern general-purpose LLMs versus specialized alignment algorithms using Russian-English parallel data. The research involved testing ten state-of-the-art models (including Gemini 3 Pro, GPT-5.2, and Claude Sonnet 4.5) using various prompting strategies (zero-shot, few-shot), alongside five baseline approaches ranging from statistical methods (fast-align, eflomal) to neural network architectures (AwesomeAlign, AccAlign, BinaryAlign). Performance was evaluated based on Precision, Recall, F-measure, and Alignment Error Rate (AER) metrics using annotated data from the Russian National Corpus. Experimental results indicated that the specialized BinaryAlign algorithm maintains the lead in overall alignment quality (F-measure 0.883, AER 0.113). However, leading LLMs, specifically Gemini 3 Pro Preview and GPT-5.2, demonstrated results surpassing those of most classic and early neural network baselines. Notably, for the most effective models, including in-context examples often reduced performance compared to the zero-shot setting. Thus, modern LLMs can serve as a reliable tool for high-quality alignment in the absence of training data, opening new perspectives for processing low-resource language pairs.

Keywords: word alignment; large language models; parallel corpora; natural language processing; machine translation; Russian-English language pair

INFORMATION ABOUT THE AUTHORS

Morozov, Dmitry A. (corresponding author)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com PhD, Junior Researcher
Makhova, Aleksandra A.	ORCID iD: 0000-0002-3323-161X. E-mail: discourse@yandex.ru Researcher
Dyachenko, Pavel V.	ORCID iD: 0000-0001-8840-9406. E-mail: pavel.v.dyachenko@gmail.com PhD, Senior Researcher
Kozerenko, Anastasia D.	ORCID iD: 0000-0003-4749-0724. E-mail: akozerenko@mail.ru PhD, Senior Researcher

Funding: Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-03-2026-305, January 16, 2026, project "Applied Research on the Implementation of Artificial Intelligence Technologies in Higher Education", project code: FSMG-2025-0086).

For citation: D. A. Morozov, A. A. Makhova, P. V. Dyachenko, and A. D. Kozerenko, "Applying large language models to Russian-English word alignment", *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 48–61, 2026.
DOI: 10.18255/1818-1015-2026-1-48-61.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 33, № 1, 2026
сайт журнала: www.mais-journal.ru

Применимость больших языковых моделей в задаче пословного выравнивания русско-английских битекстов

Д. А. Морозов[1], А. А. Махова[2], П. В. Дяченко[2], А. Д. Козеренко[3]

DOI: 10.18255/1818-1015-2026-1-48-61

[1]Новосибирский государственный университет, Новосибирск, Россия
[2]Московский физико-технический институт, Долгопрудный, Россия
[3]Институт русского языка им. В.В. Виноградова РАН, Москва, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 5 февраля 2026 г.
После доработки 22 февраля 2026 г.
Принята к публикации 24 февраля 2026 г.

В настоящей статье исследуется задача автоматического пословного выравнивания параллельных текстов, являющаяся фундаментальным этапом для обучения систем машинного перевода, сопоставительного исследования языков и создания лингвистических ресурсов. В условиях дефицита аннотированных данных для многих языковых пар особую актуальность приобретает вопрос применимости больших языковых моделей (LLM), обладающих высокими обобщающими способностями и способных решать многие задачи без длительного обучения на целевой выборке. Работа посвящена сравнительному анализу эффективности современных LLM общего назначения и специализированных алгоритмов выравнивания на материале русско-английской языковой пары. Проведённое исследование включало тестирование десяти передовых моделей (в том числе Gemini 3 Pro, GPT-5.2, Claude Sonnet 4.5) с использованием различных стратегий промптирования (zero-shot, few-shot), а также пяти базовых подходов: от статистических методов (fast-align, eflomal) до нейросетевых архитектур (AwesomeAlign, AccAlign, BinaryAlign). Оценка качества производилась на основе метрик точности, полноты, F-меры и AER с использованием размеченных данных Национального корпуса русского языка. Результаты экспериментов показали, что специализированный алгоритм BinaryAlign сохраняет лидерство по совокупному качеству разметки (F-мера 0.883, AER 0.113). Однако ведущие LLM, в частности Gemini 3 Pro Preview и GPT-5.2, продемонстрировали результаты, превзошедшие большинство классических и ранних нейросетевых решений. Примечательно, что для наиболее эффективных моделей добавление примеров в контекст часто снижало качество по сравнению с режимом zero-shot. Таким образом, современные LLM могут служить надежным инструментом для высокоуровневого выравнивания в условиях отсутствия обучающих выборок, что открывает новые перспективы для обработки малоресурсных языковых пар.

Ключевые слова: пословное выравнивание; большие языковые модели; параллельные корпусы; обработка естественного языка; машинный перевод; русско-английская языковая пара

ИНФОРМАЦИЯ ОБ АВТОРАХ

Морозов, Дмитрий Алексеевич (автор для корреспонденции)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Канд. тех. наук, младший научный сотрудник
Махова, Александра Александровна	ORCID iD: 0000-0002-3323-161X. E-mail: discourse@yandex.ru Научный сотрудник
Дяченко, Павел Владимирович	ORCID iD: 0000-0001-8840-9406. E-mail: pavel.v.dyachenko@gmail.com Канд. тех. наук, старший научный сотрудник
Козеренко, Анастасия Дмитриевна	ORCID iD: . E-mail: akozerenko@mail.ru Канд. филол. наук, старший научный сотрудник

Финансирование: Министерство науки и высшего образования Российской Федерации (соглашение № 075-03-2026-305 от 16 января 2026 г., проект «Прикладные исследования по внедрению технологий искусственного интеллекта в высшее образование», шифр: FSMG-2025-0086).

Для цитирования: D. A. Morozov, A. A. Makhova, P. V. Dyachenko, and A. D. Kozerenko, “Applying large language models to Russian-English word alignment”, *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 48–61, 2026.
DOI: 10.18255/1818-1015-2026-1-48-61.

Введение

Одним из ключевых инструментов современной лингвистики являются корпусы — обогащённые лингвистической разметкой коллекции текстов, представляющих какой-либо языковой домен или язык в целом. Появление корпусов и переход от интуитивных суждений к анализу репрезентативных массивов текстов позволили сделать изучение языка более объективным. Кроме того, накопление больших коллекций текстов дало решающий импульс развитию методов обработки естественного языка, что привело в конечном счёте к появлению современных нейросетевых инструментов, включая большие языковые модели (англ. *Large Language Model, LLM*), способных во множестве задач обработки естественного языка достигать качества, сопоставимого с экспертным, или превосходить его.

Частным случаем корпусов являются параллельные корпусы. Их основным отличием является наличие двух или более версий одного и того же произведения, как правило, на разных языках. При этом каждая из версий разбита на фрагменты, между которыми устанавливается семантическое соответствие (*выравнивание*). Размеры фрагментов определяются исходя из назначения корпуса и обычно равны одному или нескольким предложениям. Параллельные корпусы используются для сопоставительного исследования языков, в рамках преподавания перевода и для обучения моделей машинного перевода [1, 2].

Однако для решения ряда высокоуровневых задач выравнивания по фрагментам оказывается недостаточно. В частности, возникает необходимость в пословном выравнивании (англ. *word alignment*) — процедуре установления связей между лексическими единицами исходного и переведённого текста. Такая гранулярная разметка используется при автоматизации лексического анализа, выявлении переводческих эквивалентов и анализе влияния контекста на перевод [3]. Несмотря на востребованность такого типа разметки, она сравнительно редко встречается в крупных корпусных проектах. Это объясняется в первую очередь трудоёмкостью ручного пословного выравнивания, так как количество соотносимых единиц на порядок выше в сравнении с классическим пофрагментным выравниванием. В связи с этим особенно актуальны автоматизированные подходы к пословному выравниванию.

Алгоритмы пословного выравнивания активно развивались в течение последних 40 лет, начиная со статистических алгоритмов [4—6] и заканчивая применением актуальных предобученных моделей на базе архитектуры Transformer [7—12]. Качество выравнивания при этом стабильно повышалось, достигнув для большинства исследованных языковых пар значения 0.1 и ниже по метрике Alignment Error Rate, AER [5] (подробнее о вычислении этой метрики см. раздел 2).

В то же время остаётся неизученным вопрос применимости в задаче пословного выравнивания передовых больших языковых моделей. Использование механизма рассуждения (англ. *reasoning*), увеличение числа параметров и объёма исходной обучающей выборки позволяет генеративным моделям последних поколений добиваться экспертного качества во многих задачах обработки естественного языка, в том числе в случае малого количества размеченных данных и невозможности полноценного дообучения. В рамках настоящей работы мы стремились ответить на вопрос: способны ли современные системы общего искусственного интеллекта, не подвергавшиеся специальному дообучению на задачах пословного выравнивания, превзойти узкоспециализированные алгоритмы, полагаясь исключительно на свои обобщающие способности и глубокое знание мировых языков?

Таким образом, в настоящей статье исследуется задача автоматического пословного выравнивания параллельных текстов, являющаяся фундаментальным этапом для обучения систем машинного перевода, сопоставительного исследования языков и создания лингвистических ресурсов. Для изучения мы выбрали русско-английскую языковую пару как наиболее актуальную для носителей русского языка, и в то же время слабо изученную. Мы сопоставили качество нескольких подходов

с использованием десяти различных LLM с качеством пяти базовых алгоритмов, от статистических методов (fast-align, eflomal) до нейросетевых архитектур (AwesomeAlign, AccAlign, BinaryAlign). Оценка качества производилась на основе метрик точности, полноты, F-меры и AER с использованием размеченных данных Национального корпуса русского языка. В результате нескольким LLM (в частности, Gemini 3 Pro Preview от Google[1] и GPT-5.2 от OpenAI[2]) удалось превзойти большинство базовых решений, однако совокупное лучшее качество выравнивания продемонстрировал алгоритм BinaryAlign (F-мера 0.883, AER 0.113). Тем не менее, полученные нами результаты позволяют утверждать, что LLM могут быть использованы в качестве высокоуровневого базового алгоритма для пословного выравнивания, в особенности при отсутствии размеченных данных для обучения. Мы также сравнили различные стратегии добавления примеров в промпт и обнаружили, что эффективность стратегии значительно зависит от конкретной LLM. Для показавших наибольшую эффективность Gemini 3 Pro Preview и GPT-5.2 лучшего результата удалось добиться без использования примеров в промпте.

Настоящая статья организована следующим образом: в разделе 1 приводится краткий обзор актуальных методов, применяемых для пословного выравнивания текстов; в разделе 2 обсуждаются метрики, используемые для оценки качества пословного выравнивания; в разделе 3 описаны протестированные подходы к пословному выравниванию на базе LLM; в разделе 4 перечислены использованные для сравнения базовые подходы; в разделе 5 описан использованный для экспериментов набор данных; раздел 6 содержит результаты проведённых экспериментов.

1. Обзор предметной области

Одними из первых алгоритмов, показавших относительно высокое качество выравнивания, оказалось семейство IBM-моделей [4]. Эти алгоритмы представляют собой множество из пяти генеративных алгоритмов возрастающей сложности, основанных на байесовском подходе к машинному переводу. В основе их работы лежит предположение, что предложение на целевом языке может быть получено преобразованием из предложения на исходном языке через скрытый процесс выравнивания. Обучение моделей происходит без учителя на наборах пар предложений из параллельных корпусов с использованием EM-алгоритма.

Эволюция от IBM-1 до IBM-5 отражает стремление авторов алгоритмов учесть более сложные лингвистические закономерности. IBM-1 является наиболее простым подходом и оперирует исключительно лексическими вероятностями перевода, игнорируя порядок слов. IBM-2 добавляет учёт абсолютных позиций слов, связывая вероятность выравнивания с длиной предложений и порядковыми номерами токенов. В IBM-3 авторы вводят понятие «фертильности» (англ. *fertility*) — способности одного слова исходного языка порождать ноль, одно или несколько слов целевого языка. IBM-4 и IBM-5 модифицируют этот подход, моделируя перестановки слов и группируя лексику в классы для более точного учёта контекста и синтаксических различий между языками.

Ключевым недостатком IBM-моделей является их асимметричность: в зависимости от версии они моделируют связи «один-к-одному» или «один-ко-многим», но не «многие-ко-многим». Это означает, что одному слову в оригинальном тексте может соответствовать несколько слов перевода, но не наоборот, что не всегда отражает языковую реальность. Для преодоления этого недостатка на практике применяется двунаправленное выравнивание с последующей эвристической симметризацией результатов. Наиболее известной реализацией данных алгоритмов является инструмент GIZA++ [5] и его более современные, оптимизированные аналоги (например, fast-align [6]), которые, несмотря на развитие нейросетевых подходов, до сих пор часто используются в качестве надежного базового решения при оценке качества выравнивания. В то же время, качество выравнивания ста-

[1] https://deepmind.google/technologies/gemini/
[2] https://openai.com/

тистическими методами достаточно низко, и для многих языковых пар располагается в диапазоне 0.25–0.5 согласно метрике AER (см. раздел 1) [7, 11].

Как и в большинстве областей компьютерной лингвистики, на сегодняшний день лидирующие позиции в задаче пословного выравнивания занимают различные нейросетевые алгоритмы, в первую очередь опирающиеся на предобученные многоязычные языковые модели. Одним из первых подобных подходов стал алгоритм SimAlign [7], извлекающий из предобученной модели контекстуализированные векторные представления слов для обоих предложений и строящий на их основе матрицу попарного сходства между токенами. Такой подход позволил повысить качество в сравнении со статистическими алгоритмами: например, для англо-чешской языковой пары AER оказалось равным 0.13 (0.18 в случае GIZA++), для англо-хинди — 0.39 (0.49 в случае GIZA++). Эта идея получила развитие в рамках алгоритмов AwesomeAlign [8] и AccAlign [9]. В AwesomeAlign исходное пространство эмбеддингов адаптируется для задачи выравнивания: языковая модель дообучается на парах предложений без пословной аннотации с использованием функции потерь, которая сближает векторы параллельных предложений и отдаляет случайные пары. В свою очередь, AccAlign базируется на гипотезе о том, что информация, необходимая для выравнивания, распределена по разным слоям нейронной сети неравномерно. Алгоритм использует механизмы накопления весов внимания из различных слоёв модели для формирования более надежной матрицы соответствий, сглаживая случайные выбросы, характерные для отдельных слоёв. Этот подход позволяет получать более стабильные результаты на языковых парах с существенными структурными различиями, где одного лишь верхнего слоя трансформерной сети часто оказывается недостаточно. Среднее значение AER при тестировании AccAlign на материале семи языковых пар с участием английского языка оказалось равным 0.16 (0.24 для SimAlign).

Ряд исследований в этой области был также направлен на преодоление проблемы фразовых соответствий, актуальной для языков, принципиально различающихся на морфосинтаксическом уровне. Авторы алгоритма SpanAlign [13] предложили переформулировать задачу выравнивания и вместо поиска связей «токен-токен» перейти к обнаружению в тексте выравниваемых фрагментов. Такой подход позволил улучшить работу с идиомами и устойчивыми словосочетаниями. Алгоритм WSPAlign [10] (от англ. *Word Structure and Position Alignment*) был разработан для компенсации недостатков чисто семантического сопоставления. Поскольку в глубоких слоях трансформерных сетей позиционная информация часто теряется, WSPAlign в явном виде интегрирует структурные и позиционные признаки в процесс выравнивания, предотвращая ошибочное связывание схожих по семантике слов, находящихся в разных синтаксических позициях. Такой подход позволил достичь AER ниже 0.1 сразу для нескольких языковых пар: китайско-английской, румынско-английской и англо-французской. В ходе дальнейшего тестирования этого подхода среднее значение AER на материале пяти языковых пар составило 0.09 [11].

Наибольшую эффективность на сегодняшний день демонстрирует алгоритм BinaryAlign [11], авторы которого предложили перейти от выравнивания матриц эмбеддингов и поиска токена, наиболее схожего с рассматриваемым, к задаче бинарной классификации, решаемой для каждой возможной пары токенов. Для этого языковая модель дообучается на размеченных вручную данных. Авторам удалось добиться лучшего качества в сравнении с предыдущими подходами на материале пяти языковых пар: немецко-английской, румынско-английской, французско-английской, китайско-английской и японско-английской. В большинстве случаев авторам удалось добиться точности и полноты генерируемой разметки более 90 %, а среднее значение AER составило 0.07.

Одной из основных трудностей для исследований в области пословного выравнивания является крайне малое количество доступных аннотированных данных. Авторы SemiAlign [12] предложили использовать модель BinaryAlign, обученную на малом числе размеченных вручную примеров, для генерации дополнительной разметки, а затем обучить новую модель на объединении двух на-

боров данных. Такой подход позволил повысить устойчивость модели. Кроме того, авторы изучили возможность использования схожего подхода для языковых моделей-декодировщиков и обнаружили, что применение сравнительно небольших моделей (8 миллиардов параметров) позволило превзойти модели-кодировщики для части исследованных языковых пар.

Альтернативным способом преодоления проблемы нехватки данных может стать использование промптируемых больших языковых моделей, способных решать многие задачи без дообучения, опираясь лишь на небольшое число примеров, переданных в промпте (few-shot). При этом исследований, направленных на тестирование подобного подхода к выравниванию, нам обнаружить не удалось. На заполнение этой лакуны и направлена настоящая работа.

2. Оценка качества выравнивания

Стандартная экспертная разметка пословного выравнивания содержит метки двух типов: отношение S (от англ. *sure*) для однозначно выравниваемых слов и отношение P (от англ. *possible*) для тех связей, которые могут существовать, а могут и не существовать в зависимости от контекста. Каждое из отношений представляет собой множество пар вида (w_1, w_2), где w_1 — слово из исходного предложения, w_2 — слово из целевого предложения. При этом предполагается, что отношение S является подмножеством P, то есть любая S-пара является одновременно и P-парой (но не наоборот). В отличие от экспертной разметки, генерируемые алгоритмами соответствия принято относить к единственному множеству A. На базе такого подхода Ох и Ней [5] предложили метрики, к настоящему времени ставшие общепринятыми для задачи пословного выравнивания: точность (P), полнота (R), F-мера (равная среднему гармоническому точности и полноты) и AER (от англ. *alignment error rate*)

$$P = \frac{|A \cap P|}{|A|};\ R = \frac{|A \cap S|}{|S|};\ AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}.$$

Таким образом, точность отражает долю корректных связей среди сгенерированных, а полнота — долю обнаруженных связей, причём в обоих случаях за счёт учёта P/S-связей вычисляется фактически оценка сверху. Предыдущие эксперименты [9—11] продемонстрировали, что абсолютные значения метрик зависят от конкретного набора данных, однако их использование позволяет эффективно соотносить качество алгоритмов между собой.

Важно отметить, что использование подобной асимметричной схемы с двумя классами в экспертной разметке и всего одним в генерируемой обусловлено техническими ограничениями ранних подходов, для которых попытка различать классы среди генерируемой разметки, скорее всего, привела бы к крайне низкому уровню качества. Так как эти метрики являются общепринятыми, мы решили использовать именно их. Тем не менее, мы предполагаем, что в дальнейшем как алгоритмы, так и метрики качества в этой задаче могут быть адаптированы к двухклассовой системе генерируемых меток.

3. Алгоритмы выравнивания на базе LLM

В ходе предварительных экспериментов мы определили четыре подхода к генерации выравнивания при помощи LLM: подход без примеров (zero-shot) и три подхода с примерами в промпте (few-shot), различающиеся стратегией обработки S- и P-связей. Пример использованного промпта приведён на рис. 1. Во всех подходах на вход модели подавались правила выравнивания и пара предложений. В ходе предварительных экспериментов мы столкнулись с различной токенизацией текста разными моделями. Для того, чтобы избежать проблем с сопоставлением генерируемой и эталонной разметки, предложения были предварительно токенизированы: каждому токену в предложении был приписан через разделитель его порядковый номер. В первом из few-shot-подходов (few-shot-flat) мы объединили S- и P-связи в единый список и подавали на вход модели по десять

```
You are a professional linguist specializing in word alignment.

Task: Map Source Index to Target Index based on semantic equivalence, distinguishing
     between confidence levels.

Categories:
1. SURE (sure): Direct translations, named entities, exact matches. Words that MUST
    be aligned.
2. POSSIBLE (possible): Fuzzy matches, idiomatic changes, function words that
    perform similar roles but aren't direct translations.

Rules:
1. Output valid JSON object: {"id": {"sure": [[s, t], ...], "possible": [[s, t],
    ...]}}
2. IMPORTANT: The JSON keys MUST be the exact ID strings provided in the input.
3. Use 0-based indexing.

Here are examples showing both SURE alignments (direct equivalents) and POSSIBLE
    alignments (contextual/fuzzy matches):

--- Example Input ID:552 ---
SRC: 0:" 1:You 2:‘ve 3:got 4:dark 5:glasses 6:? 7:" 8:he 9:asked 10:.
TGT: 0:- 1:У 2:вас 3:есть 4:темные 5:очки 6:? 7:- 8:спросил 9:он 10:.
--- Example Output ---
{"552": {"sure": [[2, 3], [3, 3], [4, 4], [5, 5], [6, 6], [9, 8], [8, 9], [10, 10]],
     "possible": [[1, 2]]}}

...

--- Input ID:1 ---
SRC: 0:But 1:at 2:the 3:moment 4:, 5:right 6:in 7:the 8:intense 9:moment 10:, 11:I
    12:‘m 13:telling 14:you 15:- 16:nobody 17:caughts 18:.
TGT: 0:А 1:вот 2:в 3:самый 4:напряженный 5:, 6:самый 7:решающий 8:момент 9:, 10:увер
    яю 11:вас 12:, 13:никто 14:не 15:кашлянет 16:.
--- Output ---
```

Fig. 1. Prompt for few-shot-2level approach

Рис. 1. Пример использованного промпта для few-shot-2level-подхода

случайных примеров выравнивания. Во втором подходе (few-shot-sure) мы исключили из примеров *P*-связи и подавали в примерах только *S*-связи (аналогично авторам алгоритма BinaryAlign [11]). В третьем подходе (few-shot-2level) мы указали в промпте необходимость генерировать два списка пар: для *S*-связей и для *P*-связей. На выходе ожидался JSON-объект, содержащий информацию о выравнивании целевой пары предложений (в виде списка или, в случае few-shot-2level, двух списков). При подсчёте качества разметки для few-shot-2level сгенерированные списки *S*- и *P*-связей объединялись.

Для доступа к LLM мы использовали API сервиса OpenRouter[3]. При выборе LLM мы постарались задействовать актуальные на январь 2026 года модели, значительно различающиеся по позициям в открытых рейтингах качества и по стоимости обработки запроса, чтобы сформировать более объективное представление об ограничениях тестируемого подхода. В экспериментах были рассмотрены следующие модели:

- Gemini 3 Pro Preview и Gemini 3 Flash Preview от Google[4];

[3] https://openrouter.ai
[4] https://deepmind.google/technologies/gemini/

- Claude Sonnet 4.5 и Claude Haiku 4.5 от Anthropic[5];
- GPT-5.2 и GPT-5-Mini от OpenAI[6];
- Grok 4 и Grok 4.1 Fast от xAI[7];
- Kimi K2.5 от MoonshotAI[8];
- DeepSeek V3.2 от DeepSeek[9].

Лишь две из десяти рассмотренных моделей являются моделями с открыто доступными весами: Kimi K2.5[10] и DeepSeek V3.2[11]. Использование проприетарных моделей является существенным ограничением для нашего исследования, однако их лидирующие позиции в рейтингах вынуждают оценивать возможности подхода именно на базе этих моделей.

В ходе предварительных экспериментов мы определили, что лучшего качества генерации удаётся добиться при значении температуры, равном 0, в частности из-за того, что при положительных значениях температуры модели гораздо чаще нарушают предписанный формат отклика. Объём допустимых рассуждений модели выбирался как максимально возможный (для моделей, имеющих подобную настройку). При этом объём генерации существенно зависел от модели: наибольший объём рассуждений генерировался моделями Grok 4, Kimi K2.5, Gemini 3 Pro Preview и GPT-5-Mini, а наименьший — моделями Gemini 3 Flash Preview и Claude Sonnet 4.5. При этом совокупный бюджет экспериментов не превысил 100 USD.

4. Базовые подходы

При выборе базовых алгоритмов мы стремились охватить как можно больше разнообразных подходов и архитектур, для того чтобы получить лучшее представление о месте LLM среди них. Кроме того, мы выбирали такие алгоритмы, которые ранее были протестированы на многих языковых парах. В итоговый список вошли пять подходов:

1. **fast-align** [6]. Этот алгоритм представляет собой репараметризацию классической IBM-2, разработанную для существенного ускорения процесса обучения без значительной потери качества. Суть обновления заключается в замене сложного позиционного выравнивания на единственную функцию с параметром «натяжения», который регулирует отклонение связей от диагонали матрицы выравнивания. Этот подход базируется на эмпирическом наблюдении, что слова в переводе, как правило, следуют в порядке, близком к исходному (то есть монотонно). Благодаря упрощению пространства параметров и применению вариационного EM-алгоритма, fast-align способен обрабатывать огромные массивы данных на порядок быстрее классических IBM, что сделало его индустриальным стандартом для предварительной обработки корпусов. В то же время он наследует недостатки IBM-2 в части работы с перестановками слов на больших дистанциях и способен обрабатывать только отношения «один-к-одному». Для экспериментов мы использовали оригинальную реализацию алгоритма[12].
2. **eflomal** [14]. Этот алгоритм близок к IBM-3 и, в отличие от fast-align, способен порождать отношения «один-ко-многим». Главным преимуществом алгоритма является использование разреженных априорных распределений, что позволяет эффективно работать с памятью и избегать переобучения на редких словах. Этот подход часто превосходит fast-align по качеству выравнивания, особенно на языковых парах с существенными структурными различиями, за счёт более гибкого учета лексических и позиционных вероятностей, хотя обычно и уступает

[5] https://www.anthropic.com/
[6] https://openai.com/
[7] https://x.ai/
[8] https://www.moonshot.cn/
[9] https://www.deepseek.com/
[10] Веса доступны на HuggingFace: https://huggingface.co/moonshotai/Kimi-K2.5
[11] Веса доступны на HuggingFace: https://huggingface.co/deepseek-ai/DeepSeek-V3.2
[12] https://github.com/clab/fast_align

ему по скорости. Для экспериментов мы использовали оригинальную реализацию алгоритма[13].

3. **AwesomeAlign** [8]. Нейросетевой алгоритм, опирающийся на многоязычную предобученную BERT-подобную модель и обучаемый без учителя. В отличие от SimAlign, этот алгоритм предполагает дообучение предобученной языковой модели. В то же время дообучение проводится на данных без пословного выравнивания, а его целью является преобразование векторного пространства модели таким образом, чтобы сблизить векторы парных предложений. Мы использовали оригинальную реализацию алгоритма[14], в качестве предобученной модели использовалась модель bert-base-multilingual-cased[15] [15].
4. **AccAlign** [9]. Нейросетевой алгоритм на базе предобученной BERT-подобной модели, обучаемый с учителем и разработанный для решения проблемы неравномерного распределения лингвистической информации в глубоких нейронных сетях. Авторы метода исходят из предпосылки, что полагаться исключительно на последний слой трансформера или усреднение всех слоев неэффективно, так как разные уровни модели кодируют различные аспекты языка (от морфологии до семантики). AccAlign реализует механизм накопления весов внимания, который агрегирует информацию из множества слоев, подавляя стохастический шум, свойственный отдельным картам внимания. Такой подход позволяет стабилизировать матрицу выравнивания и достигать высокой точности даже на языковых парах с сильными структурными расхождениями, где стандартные методы часто теряют контекстуальные связи. Для экспериментов мы использовали оригинальную реализацию алгоритма[16], в качестве предобученной модели использовалась модель LaBSE[17] [16].
5. **BinaryAlign** [11]. Наиболее актуальный нейросетевой алгоритм на базе предобученной BERT-подобной модели, обучаемый с учителем. В отличие от предыдущих подходов, в BinaryAlign задача пословного выравнивания рассматривается как задача бинарной классификации, где для каждой возможной пары токенов решается вопрос о том, могут ли они быть выравнены. Алгоритм сравнительно нетребователен к размеру обучающей выборки, что делает его универсальным высокоуровневым базовым решением при наличии хотя бы какого-то количества примеров, размеченных вручную. Мы использовали оригинальную реализацию алгоритма[18], в качестве предобученной модели использовалась модель XLM-RoBERTa-large [17]. Как и авторы оригинальной статьи, при обучении мы не учитывали *P*-связи.

Алгоритмы fast-align, eflomal и AwesomeAlign обучаются без учителя. Для их обучения мы подготовили выборку из 250 тысяч пар предложений на русском и английском языках из состава русско-английского параллельного подкорпуса Национального корпуса русского языка (НКРЯ) [18]. Контексты извлекались случайным образом с использованием API.

5. Данные

Тестирование алгоритмов на материале русско-английской языковой пары затруднено малым количеством открыто доступных аннотированных наборов данных. В нашей работе мы использовали доступный на платформе HuggingFace набор данных[19], подготовленный исследователями НКРЯ. Этот набор состоит из 881 пары предложений на русском и английском языках. Источником предложений послужил русско-английский параллельный подкорпус в составе НКРЯ. Набор

[13] https://github.com/robertostling/eflomal
[14] https://github.com/neulab/awesome-align
[15] https://huggingface.co/google-bert/bert-base-multilingual-cased
[16] https://github.com/sufenlp/AccAlign
[17] https://huggingface.co/sentence-transformers/LaBSE
[18] https://github.com/ubisoft/ubisoft-laforge-binaryalign
[19] https://huggingface.co/datasets/ruscorpora/rnc_enru_aligned

данных был вручную выровнен пословно с привлечением трёх разметчиков, владеющих русским языком на уровне родного и английским на уровне не ниже B2. При разметке использован принцип выравнивания с двумя типами меток, описанный ранее в разделе 2. Авторы следовали протоколу разметки, предложенному для русско-китайской языковой пары [19], дополнив его рядом правил, специфичных для конкретной языковой пары (например, в отношении выравнивания словосочетаний и глаголов с управлением). Мы использовали разбиение на обучающую и тестовую выборки из оригинального датасета (в отношении 90:10). Отметим, что сравнительно малый объём доступного набора данных в рамках задачи пословного выравнивания не является чем-то необычным и вполне сопоставим с аналогичными наборами для других языковых пар: характерный размер наборов данных, используемых в работах [9—12], так же исчисляется сотнями предложений. Это ещё раз подчёркивает актуальность использования подходов, не требующих большого объёма размеченной выборки для обучения или дообучения.

6. Результаты и их обсуждение

Полученные в ходе экспериментов результаты приведены в таблице 1. Лучший достигнутый результат согласно каждой из метрик выделен серым. Результаты, уступившие лучшему не более, чем на 5 %, выделены **жирным**. Строки, соответствующие базовым решениям, выделены синим. Результаты упорядочены по возрастанию метрики AER.

Подходам на базе LLM удалось превзойти четыре из пяти базовых решений. Тем не менее, лучшего результата согласно трём метрикам из четырёх удалось добиться при помощи алгоритма BinaryAlign (F-мера 0.883, AER 0.113), причём по метрике AER отрыв от лучшего LLM-подхода составляет более 0.03. Стоит отметить, что полученная нами оценка качества базовых решений для русско-английской языковой пары хорошо соотносится с полученными ранее результатами для других языковых пар (как в абсолютных значениях, так и относительно друг друга) [9—11].

Среди LLM-подходов лучшего результата удалось достичь в рамках zero-shot-тестирования модели Gemini 3 Pro Preview (F-мера 0.851, AER 0.146). Продемонстрированный этой моделью результат обошёл четыре из пяти базовых решений (однако разница результатов с AccAlign крайне мала). Среди прочих LLM относительно высокие результаты показали GPT-5.2 (в zero-shot-тестировании F-мера 0.809, AER 0.177) и Gemini 3 Flash Preview (в zero-shot-тестировании F-мера 0.799, AER 0.207). Тем не менее, эти модели уступили всем трём нейросетевым базовым решениям (хотя в случае GPT-5.2 разница с AwesomeAlign достаточно мала). Эти три модели в различных постановках опередили все другие LLM-подходы. Худшие результаты продемонстрировали Claude Haiku 4.5 и DeepSeek V3.2: этим моделям не удалось ни разу превзойти ни одно из базовых решений ни по одной из метрик.

Многие LLM-подходы показали крайне несбалансированные результаты: высокую точность при низкой полноте или наоборот. При этом наибольший дисбаланс продемонстрировал few-shot-подход с использованием только *S*-связей в примерах, в случае которого разрыв между точностью и полнотой составлял 0.2 и более в пользу первой метрики. Это позволило подобной стратегии ожидаемо показать высокую в сравнении с прочими подходами точность, что может быть использовано в сценариях, для которых точность разметки важнее её полноты. Стоит отметить, что схожим перекосом в сторону точности обладает и BinaryAlign (тоже обученный только на *S*-связях).

В целом, влияние примеров, передаваемых в промпте, разделило модели на два типа. В случае Gemini 3 Pro Preview, GPT-5.2, Gemini 3 Flash Preview, Grok 4 и Grok 4.1 Fast использование примеров снижало качество генерируемой разметки, то есть лучших результатов удалось добиться в рамках zero-shot. Для большинства остальных моделей ситуация обратная: zero-shot-подход оказывался в числе худших.

Лучшего качества при двухуровневой разметке удалось добиться при помощи Gemini 3 Pro Preview, однако качество разметки в этом сценарии оказалось хуже всех нейросетевых базовых решений. Относительно низкое качество, достигнутое моделями в ходе few-shot-тестирования с ис-

Table 1. Experimental results **Таблица 1.** Результаты экспериментов

Подход	Точность	Полнота	F-мера	AER
BinaryAlign	0.927	**0.842**	0.883	0.113
Gemini 3 Pro — zero-shot	0.866	**0.836**	**0.851**	0.146
AccAlign	0.851	0.855	**0.853**	0.147
Gemini 3 Pro — few-shot-sure	**0.897**	0.772	0.830	0.157
Gemini 3 Pro — few-shot-flat	0.832	**0.820**	0.826	0.173
AwesomeAlign	0.811	**0.845**	0.828	0.176
GPT-5.2 — zero-shot	0.879	0.749	0.809	0.177
Gemini 3 Pro — few-shot-2level	0.839	0.788	0.813	0.181
GPT-5.2 — few-shot-flat	0.858	0.738	0.793	0.193
GPT-5.2 — few-shot-sure	**0.890**	0.694	0.780	0.198
Gemini 3 Flash — zero-shot	0.771	**0.830**	0.799	0.207
Gemini 3 Flash — few-shot-flat	0.801	0.762	0.781	0.215
GPT-5.2 — few-shot-2level	0.838	0.679	0.750	0.232
Claude Sonnet 4.5 — few-shot-2level	0.755	0.780	0.767	0.235
Claude Sonnet 4.5 — few-shot-sure	0.827	0.675	0.743	0.244
Gemini 3 Flash — few-shot-2level	0.740	0.775	0.757	0.247
Grok 4.1 Fast — few-shot-sure	0.876	0.623	0.728	0.251
Gemini 3 Flash — few-shot-sure	0.832	0.640	0.724	0.259
Kimi K2.5 — few-shot-flat	0.729	0.734	0.732	0.269
Claude Sonnet 4.5 — few-shot-flat	0.729	0.726	0.728	0.272
GPT-5-Mini — few-shot-sure	0.783	0.658	0.715	0.273
GPT-5-Mini — few-shot-flat	0.720	0.688	0.704	0.293
Grok 4.1 Fast — zero-shot	0.841	0.567	0.678	0.299
Grok 4 — zero-shot	0.661	0.758	0.706	0.304
Grok 4.1 Fast — few-shot-flat	0.733	0.614	0.669	0.319
Grok 4 — few-shot-sure	0.710	0.618	0.661	0.331
Grok 4.1 Fast — few-shot-2level	0.709	0.614	0.658	0.334
eflomal	0.596	0.759	0.668	0.346
Kimi K2.5 — zero-shot	0.643	0.631	0.637	0.361
GPT-5-Mini — few-shot-2level	0.617	0.611	0.614	0.385
Claude Sonnet 4.5 — zero-shot	0.619	0.582	0.600	0.396
fast-align	0.530	0.725	0.613	0.399
Grok 4 — few-shot-flat	0.541	0.677	0.602	0.414
GPT-5-Mini — zero-shot	0.571	0.575	0.573	0.427
Kimi K2.5 — few-shot-sure	0.564	0.464	0.509	0.481
Grok 4 — few-shot-2level	0.428	0.667	0.521	0.504
Kimi K2.5 — few-shot-2level	0.482	0.519	0.500	0.505
Claude Haiku 4.5 — few-shot-sure	0.557	0.391	0.460	0.525
DeepSeek V3.2 — few-shot-sure	0.457	0.410	0.433	0.562
Claude Haiku 4.5 — few-shot-flat	0.377	0.441	0.406	0.600
Claude Haiku 4.5 — few-shot-2level	0.347	0.448	0.391	0.619
Claude Haiku 4.5 — zero-shot	0.323	0.383	0.350	0.656
DeepSeek V3.2 — few-shot-flat	0.300	0.367	0.330	0.677
DeepSeek V3.2 — few-shot-2level	0.278	0.378	0.320	0.690
DeepSeek V3.2 — zero-shot	0.255	0.309	0.280	0.726

пользованием двухуровневой генерируемой разметки, показывает необходимость дополнительного исследования задачи в такой постановке. Вероятным решением может стать двухэтапная генерация, в ходе которой на первом этапе устанавливаются связи, а на втором этапе происходит их классификация.

В целом, следует отметить, что хотя LLM-подходам не удалось превзойти алгоритм BinaryAlign, модели Gemini 3 Pro Preview и GPT-5.2 можно считать высокоуровневым базовым решением в рамках этой задачи. Тот факт, что лучшего качества удалось добиться без использования размеченных экспертами данных, открывает широкие возможности для использования этого подхода при пословном выравнивании в случае языковых пар, не имеющих аннотированных выборок, к которым относится и большинство пар с участием русского языка.

В ходе дополнительного тестирования моделей Gemini 3 Pro Preview и GPT-5.2 мы исследовали возможность объединения аннотируемых пар в группы (батчи). Тестирование не выявило статистически значимых изменений качества сегментации при размере группы до 20 пар, что позволяет на порядок уменьшить число запросов к LLM и значительно снизить затраты времени и ресурсов на разметку.

Заключение

В настоящей работе проведено сравнительное исследование эффективности современных LLM и специализированных алгоритмов в задаче пословного выравнивания русско-английских параллельных текстов. Эксперименты показали, что, несмотря на стремительное развитие генеративного искусственного интеллекта, специализированный подход BinaryAlign по-прежнему демонстрирует наилучшие показатели качества (F-мера 0.883, AER 0.113). Тем не менее, ведущие LLM, такие как Gemini 3 Pro Preview и GPT-5.2, показали результаты, превосходящие большинство классических и ранних нейросетевых базовых решений (fast-align, eflomal, AwesomeAlign, AccAlign). Важным наблюдением стало то, что для показавших лучшие результаты моделей стратегия промптирования zero-shot (без примеров) оказалась эффективнее, чем few-shot. Полученные результаты позволяют считать, что LLM-подходы могут быть востребованы в задаче пословного выравнивания для языковых пар с отсутствующей экспертной разметкой, тогда как при наличии качественной обучающей выборки предпочтение следует отдавать алгоритмам с дообучением типа BinaryAlign. Дальнейшие исследования в этой области могут быть направлены на адаптацию LLM-подходов для двухуровневой разметки с учётом *S*- и *P*-связей и изучение гибридных подходов, объединяющих LLM-подход с прочими алгоритмами. Ключевым ограничением глубины проведённого исследования следует считать малый размер использованной тестовой выборки в связи с небольшим объёмом публично доступных аннотированных данных.

References

[1] M. Baker, “Corpora in translation studies: An overview and some suggestions for future research”, *Target: International Journal of Translation Studies*, vol. 7, no. 2, pp. 223–243, 1995. DOI: 10.1075/target.7.2.03bak.

[2] R. Dabre, C. Chu, and A. Kunchukuttan, “A survey of multilingual neural machine translation”, *ACM Computing Surveys*, vol. 53, no. 5, p. 99, 2020. DOI: 10.1145/3406095.

[3] L. Tian, D. F. Wong, L. S. Chao, and F. Oliveira, “A relationship: Word alignment, phrase table, and translation quality”, *The Scientific World Journal*, vol. 2014, no. 1, p. 438 106, 2014. DOI: 10.1155/2014/438106.

[4] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[5] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003. DOI: 10.1162/089120103321337421.

[6] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 644–648.

[7] M. Jalili Sabet, P. Dufter, F. Yvon, and H. Schütze, "SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1627–1643. DOI: 10.18653/v1/2020.findings-emnlp.147.

[8] Z.-Y. Dou and G. Neubig, "Word alignment by fine-tuning embeddings on parallel corpora", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2112–2128. DOI: 10.18653/v1/2021.eacl-main.181.

[9] W. Wang, G. Chen, H. Wang, Y. Han, and Y. Chen, "Multilingual sentence transformer as a multilingual word aligner", in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 2952–2963. DOI: 10.18653/v1/2022.findings-emnlp.215.

[10] Q. Wu, M. Nagata, and Y. Tsuruoka, "WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 11 084–11 099. DOI: 10.18653/v1/2023.acl-long.621.

[11] G. Latouche, M.-A. Carbonneau, and B. Swanson, "BinaryAlign: Word alignment as binary sequence labeling", in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 10 277–10 288. DOI: 10.18653/v1/2024.acl-long.553.

[12] Z. Miao, Q. Wu, M. Nagata, and Y. Tsuruoka, "Improving word alignment using semi-supervised learning", in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 19 871–19 888. DOI: 10.18653/v1/2025.findings-acl.1020.

[13] K. Chousa, M. Nagata, and M. Nishino, "SpanAlign: Sentence alignment method based on cross-language span prediction and ILP", in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4750–4761. DOI: 10.18653/v1/2020.coling-main.418.

[14] R. Östling and J. Tiedemann, "Efficient word alignment with Markov Chain Monte Carlo", *Prague Bulletin of Mathematical Linguistics*, vol. 106, pp. 125–146, 2016. DOI: 10.1515/pralin-2016-0013.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891. DOI: 10.18653/v1/2022.acl-long.62.

[17] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[18] S. O. Savchuk *et al.*, "Russian National Corpus 2.0: New opportunities and development prospects", *Voprosy Jazykoznanija*, no. 2, pp. 7–34, 2024. DOI: 10.31857/0373-658x.2024.2.7-34.

[19] A. Politova, O. Bonetskaya, D. Dolgov, M. Frolova, and A. Pyrkova, "Word alignment in the Russian-Chinese parallel corpus", in *Corpus Use in Cross-linguistic Research: Paving the way for teaching, translation and professional communication*, ser. Studies in Corpus Linguistics, vol. 113, 2023, pp. 195–215. DOI: 10.1075/scl.113.11pol.