DOI 10.37882/2223-2966.2025.06-3.34

ПОДХОД ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА МЕЖДУ РУССКИМ И АРАБСКИМ ЯЗЫКАМИ С ПРИМЕНЕНИЕМ МНОГОЯЗЫЧНОЙ ОНТОЛОГИИ СОВМЕСТНО С МЕТАГРАФАМИ

AN APPROACH TO IMPROVE THE QUALITY OF MACHINE TRANSLATION BETWEEN RUSSIAN AND ARABIC USING A MULTILINGUAL ONTOLOGY TOGETHER WITH METAGRAPHS

L. Saqqour
Yu. Gapanyuk
G. Afanasyev

Summary. This paper presents a conceptual architecture that integrates multilingual ontologies and metagraph-based semantic modeling into neural machine translation systems. The proposed approach aims to enhance translation accuracy between structurally divergent language pairs, such as Russian and Arabic, by addressing lexical ambiguity, cultural nuances, and domain-specific terminology. A key innovation lies in using a semantically annotated metagraph as an intermediate representation, which supports deeper context modeling and improves alignment between source and target texts. The architecture includes stages for corpus annotation, ontology alignment (using SUMO and BabelNet), metagraph construction, and integration with Transformer-based translation models. Although the system is currently in the design and prototyping phase, the theoretical analysis and modular framework offer a promising direction for building more interpretable and semantically aware translation systems.

Keywords: machine translation, metagraph, multilingual ontology, semantic annotation, Russian-Arabic translation, semantic modeling.

Саккур Лама
Аспирант, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
lama.saqqour@gmail.com
Гапанюк Юрий Евгеньевич
К.т.н., доцент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
sfm2007@yandex.ru
Афанасьев Геннадий Иванович
К.т.н., доцент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)
gaipcs@bmstu.ru

Аннотация. В статье представлена концептуальная архитектура, объединяющая многоязычные онтологии и семантическое моделирование на основе метаграфов в рамках нейросетевого машинного перевода. Предложенный подход направлен на повышение точности перевода между языками с высокой структурной разницей, такими как русский и арабский. Он позволяет учитывать омонимию, культурные особенности и профессиональную терминологию. Ключевая инновация заключается в использовании семантически аннотированного метаграфа в качестве промежуточного представления, что способствует более глубокому моделированию контекста и улучшению соответствия между исходным и целевым текстом. Архитектура включает этапы аннотирования корпуса, онтологического сопоставления (на базе SUMO и BabelNet), построения метаграфа и интеграции с моделью перевода на архитектуре Transformer. Несмотря на текущий этап проектирования, проведённый теоретический анализ подтверждает потенциал предложенного подхода для создания более интерпретируемых и семантически точных систем перевода.

Ключевые слова: машинный перевод, метаграф, многоязычная онтология, семантическая аннотация, перевод с русского на арабский, семантическое моделирование.

Введение

В современном глобализированном мире перевод между языками играет ключевую роль в обеспечении эффективной коммуникации, будь то в бизнесе, образовании или повседневной жизни. Несмотря на значительные достижения в области нейросетей машинного перевода, перевод между языками с различной структурой, такими как русский и арабский, всё ещё вызывает серьёзные трудности [9], [10].

Арабский язык, будучи семитским по происхождению, характеризуется корневой морфологией, где один корень может создавать десятки словоформ в зависимости от модели и контекста употребления. Отсутствие огласовок в письменной речи делает автоматическое

определение смысла слова крайне затруднительным для систем машинного перевода [8]. Русский язык, в свою очередь, имеет богатую словоизменительную систему, а свободный порядок слов в предложении также затрудняет автоматический анализ [5].

Дополнительной сложностью является культурная дистанция между языками. Оба языка обладают уникальной идиоматикой, пословицами и терминологией, которые не имеют прямых эквивалентов. В результате машинный перевод часто представляет собой ограниченные: выдают синтаксически правильный, но семантически и культурно неполноценный результат.

В данной статье предлагается решение данной проблемы через внедрение промежуточного семантического слоя в виде многоязычных онтологий и метаграфа. Такая структура позволяет моделировать понятия и их связи независимо от конкретного языка, способствуя более глубокому смысловому соответствию при переводе.

Обзор литературы и технологий

Машинный перевод прошёл несколько стадий эволюции, каждая из которых базировалась на различных языках и определённые принципы обработки языка и данных. В зависимости от используемой методики данный вид технологии можно разделить на следующие основные тип систем:

1. Системы, основанные на правилах (RBMT)

Первые системы машинного перевода строились на вручную прописанных грамматических и лексических правилах. Они использовали морфологический, синтаксический и семантический анализ исходного и целевого языков. Такие решения отличались хорошей интерпретируемостью и высокой точностью в узкой предметной области, но плохо масштабировались, требовали больших трудозатрат и зависели от полноты словаря.

2. Статистические системы (SMT)

В 1990-х годах появились статистические подходы, использующие параллельные тексты для расчёта вероятностей перевода слов и фраз. Не смотря на успехи, модели SMT плохо работали на языках со сложной морфологией или свободным порядком слов, например, русском или арабском. Кроме того, они часто ошибаются в случае лексической неоднозначности.

3. Нейросетевые системы (NMT)

Современные системы машинного перевода в основном используют глубокое обучение. NMT позволяет учитывать контекст всего предложения, а не только отдельных слов, поэтому качество перевода существенно возрастает. Однако такие системы требуют больших объёмов параллельных данных и часто страдают от недостаточной интерпретируемости. Особенно это проявляется при переводе между лингвистически далекими языками, когда точность семантического соответствия снижается.

4. Гибридные системы

В попытке объединить сильные стороны предыдущих подходов появились гибридные системы, которые комбинируют нейросетевые модели с дополнительными алгоритмами или знаниями. Например:

- Яндекс.Переводчик использует гибридный подход, сочетая нейросетевой движок с алгоритмами машинного обучения на основе градиентного бустинга (CatBoost), что позволяет учитывать частотные закономерности и улучшать обработку редких слов и фраз [4].
- Systran — один из старейших игроков в области машинного перевода, и в последние годы они внедрили гибридную нейросетевую архитектуру, которая объединяет: NMT (Neural Machine Translation), лингвистические ресурсы, включая морфологические правила и словари, постредактирование на основе правил, что особенно важно для корпоративных клиентов, для повышения качества перевода в технических и юридических текстах [15].
- ABBYY Compreno Translator — одна из немногих коммерческих систем, использующих онтологически обоснованный синтаксико-семантический анализ на уровне смысловых графов. В отличие от статистических или классических правил-ориентированных подходов, система реализует интерлингвальную модель: исходный текст преобразуется в нейтральное концептуальное представление на основе онтологий и логических связей между концептами. На следующем этапе осуществляется обратное порождение (generation) текста на целевом языке, уже на основе этой абстрактной смысловой структуры. Такой подход обеспечивает высокую точность перевода, особенно в терминологически насыщенных и специализированных областях [11].

Эти примеры, которые демонстрируют, как гибридные подходы могут помочь нам преодолеть вышеуказанные ограничения. Важным элементом является то, что в таких систем можно подключать онтологий и графов знаний, тем самым добавляя смысл, который невозможно выразить, используя просто последовательность слов.

5. Место предлагаемого подхода

Настоящее исследование продолжает развитие гибридных систем, предлагая интеграцию нейросетевого

перевода с многоязычной онтологией, представленной в виде метаграфа. В отличие от традиционных деревообразных структур, метаграф позволяет учитывать более сложные связи между концептами, включая перекрёстные контекстуальные зависимости. Такое решение направлено на устранение семантической неоднозначности и повышение точности перевода между русским и арабским языками, особенно в условиях дефицита качественных параллельных корпусов.

Методология

Для преодоления ограничений традиционных систем машинного перевода, особенно в условиях языков с высокой структурной и культурной неоднородностью, таких как арабский и русский, в данной работе предлагается гибридный подход. Он предполагает внедрение семантического промежуточного слоя между исходным предложением и нейросетевым механизмом перевода. Этот слой реализуется с использованием иерархической аннотируемой метаграфовой модели, которая агрегирует данные из двух онтологий — BabelNet и SUMO — и позволяет формализовать значение предложения в виде абстрактных концептов и отношений.

А. Архитектура предлагаемой системы

Общий алгоритм работы системы включает следующие этапы:

1. Анализ исходного предложения: Предложение подаётся на вход системе, где проходит предварительный лингвистический анализ (токенизация, POS-теггинг и лемматизация).
2. Параллельный запрос к онтологиям: Концепты извлекаются из двух многоязычных онтологий. Эти онтологии служат источниками базовых семантических единиц и отношений.
3. Построение метаграфа: используя аннотированную модель метаграфа [1, 2], понятия из двух онтологий объединяются в иерархически структурированный граф, где вершины, мета-вершины и мета-эджы задают атрибуты (лексические, синтаксические, семантические), связи между понятиями моделируют процессы и отношения внутри предложения.
4. Активация агентов метаграфа: агенты функций трансформируют графы на основе абстрактных функций, и агенты правил применяют эвристики для устранения неоднозначности и генерации недостающих связей.
5. Подача семантического представления в нейросетевую модель перевода: после трансформации смыслового графа, его абстрактная форма используется как дополнительный входной сигнал для нейросетевой архитектуры (например, вектор внимания или эмбеддинги семантических ролей). Это позволяет нейросети учитывать не только линейный контекст, но и иерархическую семантику, улучшая интерпретацию сложных структур.

Б. Выбор онтологий и построение метаграфа

Чтобы добиться качественного перевода между арабским и русским языками, особенно в части передачи смысла, важно использовать подходящие онтологии — то есть структуры, описывающие значения и взаимосвязи понятий. В данной работе были выбрали две такие онтологии: BabelNet и SUMO.

BabelNet [3] — это обширная многоязычная лексическая база, которая помогает учитывать не только словарные значения, но и устойчивые выражения, а также культурные особенности речи. SUMO [14], в свою очередь, ориентирована на строгую логическую структуру понятий и формализованные связи между ними. Она помогает описывать смысл высказываний с точки зрения абстрактных категорий и логических отношений [12, 13].

Далее извлечённые из этих двух онтологий концепты объединяются в метаграф — особую структуру, которая позволяет представить смысл предложения с учётом и словарных, и концептуальных связей. Была использована модель метаграфа, основанную на подходе [1,2], которая позволяет точно учитывать контекст и неоднозначности благодаря использованию специальных агентов преобразования. Такой метаграф становится промежуточным смысловым представлением, которое затем передаётся в нейросетевую модель перевода. Это позволяет существенно повысить точность перевода и сохранить важные культурные и контекстуальные нюансы оригинального текста.

В. Создание метаграфа

На этом этапе формализуется исходный текст в виде графа, который можно аннотировать. Концепции, полученные из онтологий BabelNet и SUMO, будут собраны в организованную иерархию. Узлы графа представляют конкретные и абстрактные понятия, а ребра показывают, как они связаны между собой.

Модель включает специальные вершины и ребра, которые помогают описывать более сложные отношения, такие как причинно-следственные связи, квантификация и другие. Каждому элементу графа добавляются аннотации, которые уточняют его роль в структуре сообщения.

Такой подход помогает создать ясное смысловое пространство, в котором видно как поверхностное значение слов, так и глубокие связи между концепциями. Этот метаграф станет основой для следующего шага — активации агентов.

Г. Активация агентов метаграфа

После того как метаграф построен и аннотирован, в работу вступают агенты, отвечающие за его «оживление» — уточнение и преобразование смысловых структур. В рамках предлагаемой системы используются два типа агентов: функциональные (agF) и правила-агенты (agR) [2].

Функциональные агенты отвечают за более гибкие, абстрактные изменения структуры. Они могут, например, обобщать понятия, перестраивать синтаксис или адаптировать выражения под нужный контекст. Если в исходном языке встречается двусмысленное выражение, такой агент может заменить его на более универсальное понятие, которое легче корректно передать на языке назначения [6, 7].

Правила-агенты работают по заранее заданным правилам — как универсальным, так и специфическим для конкретной культуры или языка. Они помогают уточнить значения, разрешить омонимию, адаптировать идиомы и сохранить прагматический оттенок высказываний. Модель поддерживает повторное использование одних и тех же элементов в разных контекстах, что делает её особенно гибкой.

Вместе эти два типа агентов превращают метаграф в активную структуру, где данные и логика тесно связаны. Это даёт системе необходимую адаптивность — она может учитывать не только формальную структуру текста, но и особенности реального речевого общения.

Итоговый активный метаграф затем передаётся на вход нейросетевой модели, где и происходит окончательное преобразование на целевой язык. Таким образом, агенты играют роль мостика между глубокой смысловой интерпретацией текста и статистическими механизмами перевода, обеспечивая при этом более точную и культурно релевантную передачу смысла.

Д. Подача семантического представления в нейросетевую модель перевода

На заключительном этапе активный метаграф превращается в структурированное семантическое представление, пригодное для обработки нейросетью. В отличие от обычных моделей, которые работают просто с последовательностью слов, наша нейросеть получает многослойную структуру, включающую:

- концепты и связи между ними, полученные из онтологий;
- контекстные аннотации от агентов;
- информацию об иерархии и типах отношений (например, род–вид, часть–целое и т.д.);
- культурные метки и прагматические функции выражений.

Благодаря этому подходу нейросеть не теряет важную смысловую информацию, как это часто бывает при работе только на уровне слов. Каждый элемент метаграфа представлен в виде вложения (embedding) с добавленными метками, отражающими его позицию, тип и контекст. Такая подача совместима с архитектурой Transformer.

Пример: если фраза содержит омонимы или идиомы, после обработки онтологиями и агентами она превращается в набор концептов с чётко определёнными связями и метками. Это помогает нейросети «понять» не просто слова, а то, что они значат в конкретном контексте, что особенно важно при работе с языками, где одна и та же форма может иметь множество значений — как, например, в арабском или русском.

В итоге нейросетевая модель здесь выступает не просто как «переводчик», а как интерпретатор уже осмысленного содержания. Такой подход повышает точность передачи значений, снижает вероятность смысловых искажений и помогает учитывать культурные особенности оригинального текста.

Реализация и экспериментальная установка

На текущем этапе реализация системы находится в стадии концептуального проектирования и подготовки экспериментальной базы. Основные компоненты архитектуры определены и описаны, а логика взаимодействия между модулями — формализована. Ниже представлены ключевые этапы, заложенные в структуру реализации.

1. Сбор и аннотирование корпуса

Планируется формирование параллельного корпуса предложений на русском и арабском языках, охватывающего как общеупотребительную лексику, так и доменные тексты с выраженной терминологической нагрузкой. Аннотирование предполагает выделение концептов, синтаксических ролей, прагматических маркеров и культурно-специфических элементов.

2. Онтологическое сопоставление: интеграция SUMO и BabelNet

Ключевым этапом является сопряжение двух онтологий — BabelNet, ориентированной на лексико-семантические связи, и SUMO, предлагающей формализованную логическую структуру понятий. Для обеспечения согласованности между этими онтологиями реализуется модуль онтологического сопоставления, в котором:

- BabelNet используется для извлечения лексических единиц, синонимических рядов и межъязыковых соответствий;

- SUMO — для идентификации логических и категориальных отношений между концептами, таких как агентивность, атрибутивность, причинность;
- Формируется таблица сопоставления (alignment), связывающая synset'ы BabelNet с концептами SUMO через промежуточные онтологические маркеры, включая части речи, роли и типы сущностей.

3. Построение метаграфа

На основе онтологически сопоставленных концептов формируется аннотируемый метаграф, согласно модели [7]. Реализация включает:

- Создание графовой структуры с иерархией узлов и связей;
- Назначение каждому элементу метаграфа атрибутов: онтологического типа (из SUMO), лексической формы (из BabelNet), синтаксической функции и контекста;
- Добавление мета-связей, отражающих абстрактные отношения (например, логические зависимости, антецеденты, модальность).

4. Интеграция с нейросетевой моделью машинного перевода

Готовый метаграф используется как семантический «мост» между оригинальным текстом и системой машинного перевода. Он позволяет не просто передавать слова, а транслировать смысл. Есть два способа, как его можно встроить в работу нейросети:

- На этапе энкодера — граф преобразуется в векторы признаков (embeddings), которые подаются в модель вместе с текстом и дополняют позиционную информацию;
- На этапе декодера — концепты из метаграфа помогают выбрать наиболее подходящие варианты перевода среди предложенных моделью (так называемый reranking).

Для реализации этой части можно использовать существующие открытые платформы, такие как OpenNMT, Marian NMT или Fairseq, с незначительными изменениями в коде, чтобы они могли обрабатывать структурированные входные данные.

5. Оценка качества перевода

Когда система будет реализована, её работу планируется оценивать как автоматически, так и с участием экспертов. Используются следующие подходы:

- BLEU — метрика, которая измеряет совпадения между машинным и эталонным переводом;
- TER (Translation Edit Rate) — показывает, сколько правок нужно внести в перевод, чтобы он стал точным;
- Экспертная оценка — когда носители языка или специалисты проверяют, насколько хорошо переданы смысл, стиль и культурные особенности оригинального текста.

Такой комплексный подход позволит оценить не только техническую точность перевода, но и его качество с точки зрения восприятия и адекватности.

Результаты и обсуждение

Результаты логического проектирования системы позволяют сделать обоснованные выводы о её потенциале для решения рассматриваемой задачи. Предлагаемый подход — интеграция онтологически аннотированного метаграфа в нейросетевую модель перевода — открывает возможности для повышения семантической точности. Предлагаемый подход — интеграция онтологически аннотированного метаграфа в нейросетевую модель перевода — открывает возможности для повышения семантической точности.

Использование метаграфа в качестве промежуточного семантического представления даёт ряд важных преимуществ:

- Повышение согласованности перевода за счёт устранения омонимии и неоднозначностей на этапе семантического анализа;
- Улучшение обработки терминологии и идиоматических выражений благодаря онтологической нормализации концептов;
- Снижение потерь при передаче контекста и прагматических значений оригинального текста;
- Повышение адаптивности перевода при смене доменов за счёт модульной структуры агентов.

Кроме того, использование многоязычных онтологий (BabelNet, SUMO) позволяет расширить охват языковых конструкций и учесть культурно-специфические особенности.

На следующем этапе планируется перейти к практической части — провести серию экспериментов с параллельным корпусом и оценить качество перевода с помощью как автоматических метрик (BLEU, TER), так и экспертных оценок. Это позволит количественно проверить гипотезу о преимуществах предложенной гибридной архитектуры, сочетающей нейросетевой и онтологический подходы.

Заключение

В данной работе был предложен подход к машинному переводу с русского на арабский язык, который опирается на многоязычные онтологии и аннотируемые метаграфы. Такая интеграция позволяет не просто пере-

водить текст, а глубже понимать его смысл. Благодаря объединению данных из BabelNet и SUMO удаётся создать структурированное семантическое представление высказывания — своего рода «каркас смысла», в котором отражены ключевые понятия, связи между ними, а также контекстуальные и культурные особенности.

Когда это представление подаётся в нейросетевую модель, перевод становится не только точнее, но и ближе к оригинальному замыслу. Термины передаются корректнее, снижается риск двусмысленности, а культурно значимые выражения интерпретируются более адекватно.

Таким образом, метаграф на основе онтологий становится важным мостом между знаниями о языке и современными технологиями перевода. Это делает перевод более осмысленным, гибким и устойчивым к типичным ошибкам, что особенно важно при работе с языками с разной грамматикой, логикой и культурным фоном.

ЛИТЕРАТУРА

1. Белянова М.А., Ревунков Г.И., Афанасьев Г.И., Гапанюк Ю.Е. Автоматическая генерация вопросов на основе текстов и графов знаний // Динамика сложных систем — XXI век. 2020. Т. 14. № 4. С. 55–64.
2. Гапанюк Ю.Е. Метаграф как основа моделирования релевантного медиадискурса // Science Journal of VolSU. Linguistics. 2024. Т. 23. № 5. doi: https://doi.org/10.15688/jvolsu2.2024.5.2
3. BabelNet. The largest multilingual encyclopedic dictionary and semantic network. [Electronic resource]. URL: https://babelnet.org/ (date of access 30.04.2025).
4. CatBoost is a high-performance open-source library for gradient boosting on decision trees. [Electronic resource] // Yandex CatBoost. URL: https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus (date of access 30.04.2025).
5. Corbett G.G. Gender and case in Russian // Journal of Linguistics. 1992. Vol. 28. No. 1. P. 113–138.
6. Gapanyuk Y.E. Metagraph Approach to the Information-Analytical Systems Development // CEUR Workshop Proceedings. 2019. Vol. 2514. P. 428–438.
7. Gapanyuk Y.E. The Metagraph Multiagent System Based on the Semantic Complex Event Processing // Procedia Computer Science. 2020. Vol. 169. P. 248–253.
8. Habash M. Introduction to Arabic Natural Language Processing // Synthesis Lectures on Human Language Technologies. 2010. Vol. 3. No. 1. P. 1–187.
9. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2021.
10. Katz J.S., Bever T.J. The linguistic structure of text and its implications for translation // Linguistics and Philosophy. 2019. Vol. 1. No. 3. P. 238–256.
11. Manicheva E., Petrova M., Kozlova E., Popova T. The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database // Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. Mumbai, India: The COLING 2012 Organizing Committee, 2012. P. 215–230.
12. Navigli R., Ponzetto S.P. BabelNet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network // Artificial Intelligence. 2012. Vol. 193. P. 217–250.
13. Pease A., Niles I. The Suggested Upper Merged Ontology: A large ontology for the semantic web and its applications // Proceedings of the AAAI Workshop on Ontologies and the Semantic Web. Edmonton, AB, Canada, 2002. P. 39–48.
14. SUMO: Suggested Upper Merged Ontology. [Electronic resource]. URL: https://www.ontologyportal.org/ (accessed: 30.04.2025).
15. Systran as a multilingual machine system / Toma P. // CEC. 1977. P. 569–581.

Журнал «Современная наука: актуальные проблемы теории и практики»