

ПОПОЛНЕНИЕ ТАКСОНОМИЙ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ДИАХРОНИЧЕСКИЙ ПОДХОД НА МАТЕРИАЛЕ АНГЛИЙСКОГО И РУССКОГО ЯЗЫКОВ

Садковский Фёдор Алексеевич (Fedor A. Sadkovskii)^{a,*}, Лукашевич Наталья

Валентиновна (Natalia V. Loukachevitch)^{a,**}

^a Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова»,
119991, Российская Федерация, г. Москва, Ленинские горы, д. 1

* e-mail: sadkovsky@iling-ran.ru

** e-mail: louk_nat@mail.ru

Аннотация

В настоящей работе исследуется потенциал больших языковых моделей для решения задачи автоматического пополнения таксономий на материале английского и русского языков. Авторами адаптирована передовая методика TaxoLLaMA, ранее продемонстрировавшая высокую эффективность для английского языка. В отличие от оригинального подхода, в данном исследовании используются диахронические версии тезаурусов WordNet и RuWordNet, что позволяет приблизить условия эксперимента к реальным сценариям пополнения таксономических ресурсов. В ходе работы проведено сопоставление производительности мультязычных моделей, а также, на материале RuWordNet, моделей, специально адаптированных для русского языка. Ключевое отличие от исходной архитектуры TaxoLLaMA заключается в применении оригинального метода отображения сгенерированных LLM-кандидатов в узлы таксономии, основанного на использовании векторных представлений модели

FastText. Эффективность дообученных моделей в рамках модифицированного подхода сравнивается с результатами закрытых коммерческих моделей, протестированных в режимах с предъявлением примеров (few-shot) и без предъявления примеров (zero-shot). Эксперименты подтвердили успешную применимость предложенного метода к русскоязычным данным и выявили существенное преимущество моделей, специализированных для русского языка. После дообучения модель YandexGPT-5-8B-Lite превзошла наилучшие ранее достигнутые результаты, показав значение средней точности (mean average precision) 55.42. Предложенная техника дообучения позволяет значительно повысить качество пополнения таксономий по сравнению с закрытыми моделями, протестированными без дополнительного обучения.

Ключевые слова: таксономия, пополнение таксономий, большие языковые модели, WordNet, RuWordNet, TaxoLLaMA, RuAdapt.

Введение

Задача автоматического пополнения таксономий является одной из фундаментальных в области компьютерной семантики. Решение этой задачи позволяет не только ускорить разработку и поддержку таких ресурсов, как тезаурус WordNet [Miller 1995], но и лежит в основе многих прикладных систем обработки естественного языка. Подходы к её решению эволюционировали от методов, основанных на извлечении текстовых шаблонов [Hearst 1992; Nashkole et al. 2012; Roller et al. 2014; Seitner et al. 2016; Aldine et al. 2018; Roller et al. 2018], к использованию дистрибутивных векторных моделей и их комбинаций [Baroni et al. 2012; Erk 2012; Ustalov et al. 2017; Nikishina et al. 2020b; Tikhomirov, Loukachevitch 2021].

С появлением больших языковых моделей (LLM) открылись новые перспективы для данного направления. Благодаря способности LLM выявлять семантические связи и генерировать релевантные термины, исследователи смогли достичь значительного прогресса в смежных

задачах извлечения гиперонимов и построения таксономий. Одним из наиболее эффективных решений последнего времени стал подход TaxoLLaMA [Moskvoretskii et al., 2026], в котором модель LLaMA-2-7b дообучается на парах «гипоним–гипероним», извлеченных из WordNet. Данный метод продемонстрировал передовые результаты, однако его применение и валидация были осуществлены преимущественно на англоязычных данных.

Несмотря на успехи TaxoLLaMA, при более глубоком анализе предложенной методологии обнаруживается ряд ограничений. Во-первых, тестирование подхода на других языках (испанском, итальянском) проводилось лишь в задаче извлечения гиперонимов, оставляя открытым вопрос о его эффективности для пополнения таксономий в условиях иной языковой системы. Во-вторых, в оригинальной работе использовался стандартный принцип разделения выборки, исключающий из обучения только сами тестовые понятия. Однако, как было показано в исследовании [Sadkovskii et al., 2025], такой подход может приводить к неоправданному завышению результатов за счёт присутствия в обучающих данных ко-гипонимов целевых слов, которые служат неявными подсказками из-за схожести контекстов употребления.

Для более строгой и реалистичной оценки необходимы экспериментальные условия, исключающие данную форму утечки данных, например, использование диахронических датасетов [Nikishina et al., 2020b], где обучающая и тестовая выборки формируются из хронологически разных версий таксономии.

Особый интерес представляет применение LLM к русскоязычным данным. Существующие работы для русского языка [Nikishina et al., 2020a, 2022] опирались на методы, основанные на векторных представлениях, и не обеспечивали качества, достаточного для полной автоматизации процесса. В то же время в последние годы появились как мощные многоязычные модели, так и специализированные решения, разработанные специально для русского языка (YandexGPT, модели серии RuAdapt [Tikhomirov, Chernyshev, 2023, 2024]).

Эффективность последних в ряде задач NLP на русском языке оказалась выше, чем у многоязычных аналогов, что делает их перспективным инструментом и для задачи пополнения таксономий.

Настоящая работа ставит своей целью исследование применимости современных больших языковых моделей для автоматического пополнения таксономий на материале русского языка в сравнении с английским, используя усовершенствованную методологию оценивания на базе диахронических датасетов.

В работе предложена адаптированная методология на основе TaxoLLaMA для работы с диахроническими версиями WordNet и RuWordNet, что позволяет приблизить условия эксперимента к реальному сценарию пополнения ресурса. В рамках адаптированной методологии проведена серия экспериментов по дообучению мультязычных LLM (размером 7–8 млрд параметров) на англоязычных данных для верификации воспроизводимости подхода на новом типе датасетов. Проведено исследование эффективности применения мультязычных и специализированных русскоязычных LLM (включая модели, адаптированные по методике RuAdapt) к задаче пополнения таксономии RuWordNet. Качество дообученных моделей сравнивается с результатами, полученными закрытыми коммерческими моделями (YandexGPT, GigaChat) в режимах zero-shot и few-shot, а также с наилучшими ранее опубликованными результатами для русского языка. По результатам тестирования выполнен качественный анализ ошибок для определения направлений дальнейшего совершенствования предложенного подхода.

1. Обзор литературы

В развитии методов решения задачи автоматического извлечения гиперонима и пополнения таксономий можно выделить несколько ключевых парадигм. Ранние подходы основывались на

использовании лексико-синтаксических шаблонов [Hearst 1992; Nashkole et al. 2012; Roller et al. 2014; Seitner et al. 2016; Aldine et al. 2018; Roller et al. 2018], фиксирующих типичные контексты встречаемости более общего и более частного понятий. Несмотря на высокую точность, такие методы страдали от низкой полноты из-за ограниченности набора шаблонов.

С развитием дистрибутивной семантики векторные представления слов стали основой для нового поколения методов. Исследователи [Baroni et al. 2012; Erk 2012; Ustalov et al. 2017; Nikishina et al. 2020b; Tikhomirov, Loukachevitch 2021] показали, что гипонимы и гиперонимы могут быть выявлены на основе операций над их векторными представлениями. Наилучшие результаты на русскоязычных данных RuWordNet были достигнуты с использованием мета-эмбедингов, объединяющих контекстную и графовую информацию [Nikishina et al., 2022]. Однако данные подходы, как правило, не позволяли полностью автоматизировать процесс пополнения таксономий из-за ограничений в обобщающей способности.

Переломным моментом стало появление предобученных языковых моделей-энкодеров, таких как BERT [Devlin et al., 2019]. Эти модели, обучаемые на задаче маскированного языкового моделирования, естественным образом подходили для заполнения пропусков в шаблонах. В работе [Hanna, Mareček, 2021] исследовалась способность BERT корректно восстанавливать гипероним в шаблонах Херста, и хотя качество уступало специализированным системам (например, CRIM [Bernier-Colborne, Barriere, 2018]), оно превосходило все не обучаемые ранее подходы. Дальнейшее развитие энкодеров привело к созданию более сложных архитектур: TaxoPrompt [Xu et al., 2022], Hypert [Yun et al., 2023], TaxoComplete [Arous et al., 2023], TEMP [Liu et al., 2025] и CoSTC [Niu et al., 2024]. Эти методы использовали различные техники — от обучения проекционных матриц до контрастивного обучения и ранжирования таксономических путей, — но все они опирались на энкодеры в качестве ядра.

В то же время активно развивалось направление генеративных моделей-декодеров (GPT [Radford et al., 2018] и их последователи). Долгое время считалось, что декодеры хуже подходят для задач, требующих оценки вероятности конкретного токена в последовательности, однако их главным преимуществом является отсутствие ограничений на длину генерируемого текста. Это позволяет, например, генерировать словосочетания или перечисления кандидатов, что принципиально важно для задачи пополнения таксономий, где ответом может быть не одно слово, а целый синсет. В работе [Liao et al., 2023] было проведено сравнение энкодеров и декодеров на задаче предсказания гиперонимов, показавшее сопоставимые результаты.

Важным шагом в применении декодеров стала работа [Nikishina et al., 2023], в которой модели T5 и Flan-T5 (архитектуры энкодер-декодер) превзошли как чистые декодеры (GPT-2, OPT), так и предыдущие подходы на энкодерах. Авторы продемонстрировали эффективность как применения инструкций на основе шаблонов Херста без предъявления примеров (zero-shot), так и с предъявлением (few-shot).

Кульминацией этого направления стал подход TaxoLLaMA [Moskvoretskii et al. 2026], в котором модель LLaMA-2-7B дообучалась на парах «гипоним–гипероним», извлеченных из WordNet. Благодаря использованию методов низкоранговой адаптации (LoRA) и 4-битного квантования, авторам удалось создать экономичную модель «все-в-одном», способную решать задачи извлечения гиперонимов, построения и пополнения таксономий. Вариант модели TaxoLLaMA-bench, обученный на очищенной от тестовых понятий выборке, установил 11 новых рекордов на 16 датасетах, подтвердив универсальность подхода. Критически важным для TaxoLLaMA стало использование не просто пары слов, а пары «гипоним — гипероним» с учетом определения гипонима, что позволило модели более эффективно усваивать таксономические знания.

Применительно к русскому языку ситуация складывается несколько иначе. Пионерская работа [Sabirova, Lukanin, 2014] была посвящена адаптации шаблонов Херста для русского языка и тестированию на данных DBPedia. Позднее, в рамках соревнования RUSSE'2020 [Nikishina et al., 2020a], был впервые предложен диахронический датасет на основе RuWordNet, и лучшие результаты были достигнуты комбинаторными методами с привлечением внешних ресурсов. В последующих исследованиях [Nikishina et al., 2022] на этом же материале тестировались различные векторные представления, и лучший результат (47.4 MAP) был получен с помощью мета-эмбедингов. При этом, за исключением модели-энкодера RuBERT [Kuratov, Arkhipov, 2019], большие языковые модели до настоящего времени не применялись для пополнения русскоязычных таксономий. Это представляется существенным пробелом, особенно с учетом появления специализированных русскоязычных моделей, таких как YandexGPT и серия RuAdapt [Tikhomirov, Chernyshev, 2023, 2024]. Последний подход заслуживает особого внимания, поскольку он включает замену токенизатора для лучшего соответствия русской морфологии и дообучение на крупных корпусах русских текстов и инструкций (DaruLM, Darumeru).

Таким образом, на момент начала данного исследования, несмотря на впечатляющие успехи TaxoLLaMA для английского языка и наличие мощных русскоязычных LLM, вопрос об эффективности генеративных моделей для пополнения таксономии RuWordNet оставался открытым. Кроме того, оставалась неисследованной способность моделей переносить таксономические знания с одного языка на другой, а также влияние более строгой методологии оценивания (исключающей не только тестовые слова, но и их ко-гипонимы) на итоговые результаты.

2. Данные

Источником данных в настоящей работе служат два диахронических датасета:

1. Русскоязычный датасет на основе RuWordNet, впервые предложенный в [Nikishina et al., 2020a] и позже расширенный в [Nikishina et al., 2020b].
2. Англоязычный датасет на основе WordNet, созданный по аналогии с предыдущим в [Nikishina et al., 2020b].

В подобных датасетах в качестве обучающей подвыборки используется более ранняя версия ресурса (RuWordNet 1.0), а в качестве тестовой — слова, добавленные в более позднюю («RuWordNet 2.0–1.0»).

Англоязычный датасет, использованный в работе, основан на одном из самых поздних срезов — версии 2.0 в сопоставлении с 3.0. Выбор мотивирован тем, что этот датасет по количеству элементов в тренировочной и тестовой выборках наиболее близок к наборам данных, использованных при обучении TaxoLLaMA и включает 82 115 синсетов для имен существительных и 1500 понятий в тестовой выборке из «1A: English». Общая статистика по использованным датасетам приводится в таблице 3.

версия	количество синсетов (имена существительные)	количество новых слов (имена существительные)
<i>WordNet 2.0 - WordNet 3.0</i>	79 689	2 620
<i>RuWordNet 1.0 - RuWordNet 2.0</i>	14 660	2 288

Таб. 3: Подвыборки диахронических датасетов на основе WordNet и RuWordNet с именами существительными

В [Nikishina et al., 2020a] было предложено разделение новых слов RuWordNet на две подвыборки — «public» (валидационная) и «private» (тестовая). Тестовая выборка использовалась для окончательной оценки качества предложенных решений и включает 1 525 синсетов (множеств синонимов); валидационная выборка содержит 763 синсета. В качестве

правильных ответов к целевым словам были извлечены не только непосредственные гиперонимы, но и гиперонимы второго порядка.

В RuWordNet, как и в WordNet, узлы графа представляют собой синсеты (множества синонимов), однако способ их организации имеет две отличительные особенности.. Названия синсетов могут быть одним словом, словосочетанием или включать перечисление слов/словосочетаний через запятую, уточняющие слова в скобках. Другая особенность обусловлена тем, что RuWordNet наследует названия синсетов от тезауруса RuТез [Loukachevitch, Dobrov, 2002], семантическая сеть которого не подразделялась на подграфы разных частей речи. Таким образом, в зависимости от семантики слова, названия синсетов для существительных могут быть не только существительными, но и глаголами («буллинг»>«оскорбить») и прилагательными («консонанс»>«благозвучный»).

В экспериментах с моделями в режиме с предъявлением примеров и без использовались те же наборы данных. Однако, поскольку модели использовались без обучения в формате диалога с использованием инструкций, потребовалось перевести датасет в диалоговый формат. Использовалась инструкция ТахоLLaMA, переведенная на русский язык:

Ты ассистент-помощник. Перечисли все возможные слова через точку с запятой. В ответе не должно быть ничего, кроме слов через точку с запятой.

гипоним: {целевое_слово}

гиперонимы:

В набор примеров вошли следующие названия синсетов из RuWordNet, наиболее далекие друг от друга по векторным представлениям:

- 1) «лори» → «приматы»;
- 2) «мкд» → «дорога (путь сообщения); кольцевая дорога; автомобильная дорога»;

- 3) «мкрн» → «место в пространстве; жилой массив; территория, участок»;
- 4) «преф» → «корпоративная ценная бумага; долевая ценная бумага; акция; эмиссионные ценные бумаги»;
- 5) «репо» → «гражданско-правовые отношения; гражданско-правовая сделка; операция (отдельное действие)»;
- 6) «рэнд» → «денежная единица; деньги»;
- 7) «тэгу» → «муниципальное образование; город; городское поселение»;
- 8) «убор» → «изделие легкой промышленности; изделие; одежда; головной убор»;
- 9) «ушат» → «кадка, вместилище»;
- 10) «фгос» → «стандарт; национальный стандарт; нормативно-техническая документация»

Примеры подавались в зафиксированном случайном порядке, одинаковом для всех экспериментальных сессий. Датасет доступен для просмотра и скачивания на сайте базы данных Hugging Face¹.

3. Метод

Метод проведения экспериментов, описанных в этом разделе, отличался от подхода TaxoLLaMA наборами данных — обучающим и тестовым.

В качестве обучающих данных использовался диахронический датасет «WordNet 2.0 - WordNet 3.0» с именами существительными. Обучающая выборка из пар гипоним–гипероним была сформирована по алгоритму из (Moskvoretskii et al. 2024c), но без предварительного удаления из графа каких-либо узлов.

В качестве тестовых данных использовались новые понятия и соответствующие им родительские синсеты из WordNet 3.0, ранее не существовавшие в версии 2.0.

Поскольку тестирование моделей на задачу пополнения таксономии предполагает возможность обращения к множеству узлов этой таксономии, при оценивании следует

¹ <https://huggingface.co/datasets/TaxoLLMs/ruwordnet-chat-few-shot>

учитывать специфику данной задачи. Так, в отличие от более общей задачи извлечения гиперонимов, в задаче пополнения таксономии гипероним для целевого слова необходимо подобрать из множества узлов данной таксономии. Однако большие языковые модели порождают последовательности токенов, которые необязательно будут составлять какие-либо из существующих узлов. Как обсуждалось выше, узлы, то есть синсеты, в WordNet кодируются специальным образом. Однако в (Moskvoretskii et al. 2024) было показано, что обучение модели на узлах в таком виде неэффективно. Следовательно, необходим метод обработки ответа, чтобы привести предсказанный набор слов $\langle w_1, \dots, w_n \rangle$ к списку синсетов $\langle s_1, \dots, s_m \rangle$. Процедуру обработки вывода модели имеет смысл проводить в несколько шагов.

Рассматриваются только первые k кандидатов. Как и в задаче предсказания гиперонимов, обсуждавшейся выше, k устанавливается равным 15. Прежде всего, необходимо удалить повторяющиеся элементы $w_i = w_j, i \neq j < k$, в противном случае значение метрики MAP будет завышено

Во-вторых, среди предсказанных элементов можно найти те, которые присутствуют в названиях синсетов, и переместить их в начало списка, сохраняя порядок. Этот шаг должен повысить точность предсказания.

Для нахождения наиболее вероятных синсетов можно использовать либо только названия, найденные на предыдущем шаге, либо попытаться также извлечь информацию из всех предсказанных слов. Ожидается, что этот шаг может повысить полноту.

В настоящей работе рассматриваются обе возможности, описанные в пункте три. Для поиска синсета были использованы предобученные векторные представления из модели FastText²: за векторное представление синсета брались усредненные векторные представления всех входящих в него лемм, как это предложено в (Nikishina et al. 2020b). Усреднение выполнялось с помощью евклидовой нормализации:

² Предобученные векторы были взяты с официального сайта <https://fasttext.cc/docs/en/crawl-vectors.html>

$$E' = \frac{\sum_i^n E_i}{\|\sum_i^n E_i\|_2}$$

где E' — новый вектор синсета, $E_1, \dots, E_n$ — векторы входящих в него лемм. Векторы предсказанных слов — только словарных или всех k — усреднялись таким же образом, и далее производился поиск k ближайших соседней по косинусному расстоянию.

Оригинальный подход TaxoLLaMA, предложенный в [Moskvoretskii et al., 2024], включает следующие компоненты:

1. подготовка выборки на основе пар гипоним–гипероним из WordNet; тестовые понятия, входящие в датасеты для тестирования, удаляются;
2. дообучение модели LLaMA-2-7b [Touvron et al., 2023] на предсказание списка гиперонимов через запятую;
3. оценка качества на датасетах для тестирования.

В настоящей работе предлагается расширение этого метода на другие данные и модели:

1. В качестве обучающего множества берутся пары гипоним–гипероним из RuWordNet 1.0.
2. Различные 7-8 млрд-ные LLM (перечислены в след. разделе) обучаются предсказывать слова через разделитель «; », поскольку этот знак не встречается в названиях синсетов.
3. Оценка качества на диахроническом датасете «RuWordNet 2.0–1.0» (выборка «pouns-private»).

Перед оцениванием необходимо сопоставить предсказанной последовательности слов названия синсетов. Для этого в работе предложена следующая процедура:

1. В начало списка переносятся последовательности слов, соответствующие существующим в таксономии названиям синсетов.
2. Синсет переводится в векторный вид с помощью L2-нормализации (по [Bollegalo, Bao, 2017]) векторных представлений всех неслужебных слов, входящих в название синсета.

3. Max-Pooling: $k = 15$ ближайших соседей из множества векторных представлений синсетов были найдены для каждого из нерелевантных кандидатов, а затем среди объединенного множества размера $k \times m$ ($m \leq 15$ — число предсказанных кандидатов) были извлечены все названия синсетов, встречающиеся более 1 раза, и упорядочены по частоте и по позиции первого появления в общем списке $\langle x_{1,1}, \dots, x_{1,k}, \dots, x_{m,1}, \dots, x_{m,k} \rangle$.

Данный метод пост-обработки использует преимущества организации RuWordNet и позволяет достигнуть более высокого итогового качества, чем использование предсказаний в чистом виде.

В качестве векторной модели использовалась модель FastText³ [Bojanowski et al., 2017].

4. Эксперименты

4.1. Метрики

В задаче пополнения таксономии используются две основные метрики: Mean Average Precision, MAP (4.1) и Mean Reciprocal Rank, MRR (4.2):

$$MAP@K = \frac{1}{N} \sum_{n=1}^N AP@K_n, \#(4.1)$$

$$AP@K = \frac{1}{R} \sum_{k=1}^K (Precision@k \cdot I[y_k = 1]),$$

$$Precision@k = \frac{p}{k},$$

$$MRR@K = \frac{1}{N} \sum_n \frac{1}{rank_i} \#(4.2)$$

Обозначения: $K = 15$ — максимальное число предсказаний; N — размер выборки; R — количество правильных ответов (гиперонимов) для целевого слова; k — индекс элемента в списке предсказаний; I — индикаторная функция; y_k — класс (верный ответ/неверный ответ) на позиции k ; p — количество релевантных элементов среди k первых предсказанных; $rank_i$ — позиция первого релевантного элемента.

³ Предобученные векторы были взяты с официального сайта <https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.ru.300.bin.gz>

Используемая в ряде работ (в частности, [Moskvoretskii et al., 2024]) модификация Scaled MRR не используется, поскольку она предназначена для выборок, включающих только непосредственные гиперонимы.

4.2. Базовые методы

В качестве базовых методов используются базовый метод на основе векторной модели FastText (*fasttext*) и лучший метод на основе мета-эмбедингов (*AAEME (words + TADW)*, усредненная сумма графовых + контекстных представлений) из [Nikishina et al., 2022].

4.3. Модели

Для того, чтобы оценить применимость других моделей для обучения согласно подходу TaхоLLaMA, были отобраны большие языковые модели типа декодеровщик с числом параметров, равным числу параметров исходной модели LLaMA-2 — 7 млрд — или близким к нему. Выбирались выложенные в открытый доступ базовые модели (англ. Base Models), не проходившие специальное дообучение на понимание инструкций. Были протестированы следующие модели:

- LLaMA-3.1-8B
- Mistral-7B-v0.1 (Jiang et al. 2023)
- Mistral-7B-v0.3
- Qwen2.5-7B (Yang et al. 2024)
- Qwen3-8B-Base
- Gemma-7b (Team Google 2024)

Для экспериментов на RuWordNet были отобраны большие языковые модели типа декодеровщик с числом параметров, приблизительно соответствующим числу параметров исходной модели LLaMA-2 — 7–8 млрд. Выбирались модели в открытом доступе, не проходившие инструктивное дообучение:

- LLaMA-2-7B-hf⁴;
- LLaMA-3.1-8B⁵;
- Mistral-7B-v0.1 [Jiang et al., 2023];
- Mistral-7B-v0.3⁶;
- Qwen2.5-7B [Yang et al., 2024];
- Qwen3-8B-Base⁷;
- Gemma-7b [Team Google, 2024].

В список также была включена оригинальная модель TaxoLLaMA: результаты ее дообучения позволяют судить о том, насколько модель способна переносить знания о таксономических отношениях с одного языка на другой.

Другая группа моделей включала базовые модели, специально предназначенные для работы с русским языком и примерно с тем же количеством параметров:

- YandexGPT-5-Lite-8B-pretrain [Яндекс, 2025]
- Модели из серии RuAdapt [Tikhomirov, Chernyshev, 2023, 2024]:
 - RuadaptQwen2.5-7B-Lite-Beta;
 - RuAdaptLLaMA-2-7b;
 - RuAdaptMistral-v0.1;
 - RuAdaptLLaMA-3.1-8B.

Для тестирования моделей с большим числом параметров в режиме с предъявлением примеров и без были выбраны два поколения модели YandexGPT — YandexGPT-4 и YandexGPT-5 (частные; число параметров неизвестно), модель GigaChat-2-Max (частная; число параметров неизвестно) и ряд моделей с открытым кодом: LLaMA-3.1-70B, LLaMA-3.3-70B, Qwen2.5-72B и Qwen3-32B. Тестирование осуществлялось в сервисе Yandex Foundation Models.

⁴ <https://huggingface.co/meta-llama/Llama-2-7b-hf>

⁵ <https://huggingface.co/meta-llama/Llama-3.1-8B>

⁶ <https://huggingface.co/mistralai/Mistral-7B-v0.3>

⁷ <https://huggingface.co/Qwen/Qwen3-8B-Base>; <https://qwenlm.github.io/blog/qwen3/>

4.3. Результаты

Результаты моделей, обученных на данных WordNet, в сравнении с базовым подходом, предложенным в (Nikishina et al. 2020b) и передовым подходом из (Nikishina et al. 2022a) приводятся в таблице 1.

Таб. 1. Сравнение качества моделей после дообучения на «WordNet 2.0 - WordNet 3.0». Знаком «*» обозначены результаты, взятые из (Nikishina et al. 2022a) (они продублированы, т.к. не связаны с методом подсчёта)

модели	словарные		все	
	MR	MA	MR	MA
	R	P	R	P
LLaMA-2-7b-hf	41.6	33.1	44.3	34.9
	0	2	2	5
	42.3	33.3	<u>45.8</u>	36.6
LLaMA-3.1-8B	8	6	<u>9</u>	3
	41.0	32.1	43.8	35.6
	4	0	5	7
Mistral-7B-v0.1	41.9	33.3	42.7	36.2
	5	9	9	4
	43.4	36.4	46.2	41.3
Mistral-7B-v0.3	7	6	0	9
		<u>35.1</u>	45.0	<u>39.7</u>
	<u>42.4</u>	<u>4</u>	4	<u>1</u>
Qwen2.5-7B		32.5	44.2	37.4
	40.3	0	1	6
Qwen3-8B-Base				
Gemma-7b				

*fasttext ⁸	—	40.0 0	—	40.0 0
*AAEME (words + TADW) ⁹	—	48.0 0	—	48.0 0

Из протестированных моделей наиболее высокого качества достигли модели семейства Qwen. Что касается сравнений с предыдущими подходами, то превзойти базовый метод на основе FastText удалось только дообученной модели Qwen2.5-7B; превысить качество наилучшего метода из (Nikishina et al. 2022a) при этом не удалось. В целом, кандидаты, сгенерированные моделями, уступают тем, которые порождаются в предыдущих работах по алгоритму на основе FastText.

Результаты моделей, обученных на данных RuWordNet¹⁰, приводятся в таблице 2¹¹. Среди мультязычных моделей, наилучшие результаты продемонстрировали модели семейства Qwen, в особенности Qwen3-8B-Base. Тем не менее, по метрике MAP даже лучшие модели не смогли превзойти базовые методы.

Модели для русского языка в среднем показали результаты выше: 45.02 против 38.12 по MRR и 32.8 против 23.21 по MAP. Все модели, прошедшие адаптацию на русский язык, показали более высокое качество, чем их оригинальные неадаптированные версии, см. таблицу 2.

Модель на основе Qwen2.5-7B оказалась лидером как по абсолютному результату, так и по приросту метрик, который дает методика адаптации. Следующей моделью по значению метрик оказалась модель на основе LLaMA-3.1-8B, а следующей моделью по приросту качества — модель на основе Mistral-7B-v0.1.

⁸ (Bojanowski et al. 2017)

⁹ (Nikishina et al. 2022a)

¹⁰ Веса дообученных моделей доступны на сайте: <https://huggingface.co/TaxoLLMs>

¹¹ Здесь и далее значения метрик умножены на 100

Среди всех моделей самого высокого результата по метрикам MRR и MAP достигла модель YandexGPT-5-Lite-8B-pretrain. Значение метрики MAP 55.42 превосходит значение 47.4 лучшего подхода из [Nikishina et al., 2022b]. Вторую позицию занимает модель RuAdapt-Qwen2.5-7B с результатом 41.69. Этот результат не смог превысить лучший предыдущий результат, однако оказался выше, чем у базового решения (41.4).

Самого низкого качества достигла дообученная на русских данных модель TaxoLLaMA. Значения метрик оказались даже ниже, чем у исходной для нее LLaMA-2-7b-hf, что говорит о неспособности модели эффективно переносить таксономические знания из WordNet на данные RuWordNet.

Табл. 2. Сравнение качества моделей после дообучения на «RuWordNet 1.0 – 2.0». Жирный шрифт — лучший результат, подчеркивание — второй после лучшего.

модели	MR R	MAP
	25.0	17.35
LLaMA-2-7b-hf	1	
	35.1	25.35
LLaMA-3.1-8B	1	
	32.9	22.32
Mistral-7B-v0.1	4	
	32.7	21.62
Mistral-7B-v0.3	3	
Qwen2.5-7B	38.2	29.43
	40.0	30.06
Qwen3-8B-Base	6	

		30.7	23.44
Gemma-7b	9		
		24.1	16.09
TaxoLLaMA	0		
		61.5	55.42
YandexGPT-5-Lite-8B-pretrain	5		
		<u>49.4</u>	41.69
RuadaptQwen2.5-7B-Lite-Beta	7		
		30.8	25.11
RuAdaptLLaMA-2-7b	2		
		42.3	34.79
RuAdaptLLaMA-3.1-8B	2		
		40.9	33.57
RuAdaptMistral-v0.1	5		
<i>fasttext</i>		–	41.40
<i>AAEME triplet loss (words</i>		–	<u>47.40</u>

Табл. 3. Сравнение результатов моделей с адаптацией на русский язык по методике RuAdapt [Tikhomirov, Chernyshev, 2023, 2024] и без.

модель	оригинал		RuAdapt	
	MR	MA	MRR	MAP
	R	P		
LLaMA-2-7b-hf	25.0	17.3	30.82 _{+5.8}	25.11 _{+7.7}
	1	5	1	6
LLaMA-3.1-8B	35.1	25.3	<u>42.32</u> _{+7.2}	<u>34.79</u> _{+9.4}

	1	5	1	4
	32.9	22.3	40.95 _{+8,0}	33.57 _{+11.}
Mistral-7B-v0.1	4	2	<u>1</u>	<u>25</u>
	38.2	29.4	49.47_{+11.}	41.69_{+12.}
Qwen2.5-7B		3	27	26

Результаты моделей, протестированных без предъявления примеров, показаны в таблице 4. Как и в предыдущем разделе, в таблицах ниже приводятся значения метрик, полученные с применением векторной модели для улучшения подбора синсетов, поскольку в этом случае он также обеспечил наивысшее качество. Сравнение с другими методами обработки можно найти в таблицах в Приложении.

Таб. 4. Результаты тестирования больших языковых моделей с инструкцией на русском языке без предъявления примеров. Для сравнения приводятся результаты лучшего подхода с обучением из предыдущих работ и лучшей модели, обученной по методу из раздела 4.3.1.

модели	MR	MA
	R	P
	48.2	36.3
YandexGPT-5	1	8
	38.7	21.0
YandexGPT-4	8	7
	<u>43.5</u>	<u>29.2</u>
GigaChat-2-Max	<u>2</u>	<u>8</u>
	34.1	21.0
Qwen2.5-72B	6	9
Qwen3-32B	39.3	28.4

	7	1
	37.3	22.6
LLaMA-3.1-70B	6	8
	34.6	24.9
LLaMA-3.3-70B	7	6
		<u>47.4</u>
*AAEME triplet loss (words) ¹²	—	<u>0</u>
	61.5	55.4
YandexGPT-5-Lite-8B-pretrain	5	2

«Фундаментальная» версия модели YandexGPT-5 достигает наивысшего результата. Модели удалось превзойти по обеим метрикам большинство моделей из предыдущего раздела, кроме своей дообученной облегченной версии YandexGPT-5-8B-Lite-pretrain, а также модели RuAdapt-Qwen2.5-7B. Тем не менее, по метрике MAP не удалось превзойти не только предыдущий метод, но и базовый метод на основе FastText из (Nikishina et al. 2020b) — 41.4.

Остальные модели по результатам оказываются примерно на одном уровне с моделями, обученными на русских данных «с нуля», и уступают всем моделям, адаптированным по методике RuAdapt, за исключением RuAdaptLLaMA-2-7b (ср. с таблицей 17).

Сопоставляя результаты моделей одной серии, можно прийти к выводу, что на результат в данной задаче влияет не столько число параметров, сколько поколение модели: результаты LLaMA-3.3-70B (с тем же числом параметров) и Qwen3-32B (с меньшим числом параметров) оказываются выше, чем результаты моделей младшего поколения LLaMA-3.1-70B и Qwen2.5-72B, что говорит о прогрессе в области оптимизации баланса между количеством весов и качеством решения задачи языкового моделирования.

В постановке с предъявлением примеров было протестировано только три модели с наилучшим результатом, достигнутым в тестировании без примеров, то есть YandexGPT-5, GigaChat-2-Max и Qwen3-32b, см. таблицу 5.

¹² (Nikishina et al. 2022a)

Результаты показывают, что предъявление примеров положительно повлияло только на метрику MRR YandexGPT-5, благодаря чему по ней эта модель превзошла RuAdapt-Qwen2.5-7B. Однако по метрике MAP результаты всех моделей ухудшились. Это связано с тем, что примеры побуждают модели генерировать меньшее количество кандидатов, из-за чего снижается полнота и возможность получить дополнительные названия синсетов с помощью векторной модели.

Таб. 5. Результаты тестирования больших языковых моделей с инструкцией на русском языке с предъявлением 10-ти примеров. Нижним индексом подписано изменение метрик по сравнению с результатом тех же моделей в тестировании без примеров.

модели	MRR	MAP
YandexGPT-5	50.46 _{+2.}	32.79 _–
	23	3.61
GigaChat-2-Max	38.20 _–	22.62 _–
	5.32	6.66
Qwen3-32B	39.18 _–	24.40 _–
	0.19	4.01
*AAEME triplet loss (words) ¹³	–	<u>47.40</u>
YandexGPT-5-Lite-8B-pretrain	61.55	55.42

Таким образом, для получения высокого качества решения задачи действительно необходимо дообучение, в результате которого модель сформирует представление о существующих наименованиях синсетов, что облегчит соотнесение предсказанных кандидатов с их списком. Тем не менее, даже без дообучения после обработки порожденной последовательности векторной моделью рассмотренные большие языковые модели продемонстрировали уровень качества, который сопоставим с качеством большинства дообученных моделей с небольшим числом параметров или даже превосходит его. Это

¹³ (Nikishina et al. 2022a)

открывает перспективы дальнейшего исследования моделей с большим числом параметров, чем 7–8 млрд.

5. Анализ ошибок

В тестировании на русском языке можно выделить 7 основных типов ошибок. Эти типы перечислены ниже и проиллюстрированы примерами.

1. Неверно определена предметная область. Слово «авуары» (первое значение по [БТС, 2003] — «средства, за счёт которых производятся платежи и погашаются обязательства») относится к сфере финансов, однако модель Qwen3-8B-Base предсказывает гиперонимы из области этнологии: «этническая общность; этнические народы; африканцы (население); население государства; нация» при правильных «денежная единица; свободно конвертируемая валюта; банковский вклад».

2. Другие семантические отношения. Примером может служить ответ модели RuAdaptQwen2.5-7B для целевого слова «пивная»: модель через корень связала это слово с алкогольным напитком, но предсказала кандидаты, подходящие для других однокоренных слов — «пивоварни» и «пива»: «предприятие пищевой промышленности; алкогольный напиток; спиртосодержащая продукция».

3. Слишком абстрактные кандидаты. К таким примерами относится, в частности, предсказание YandexGPT-5-Lite для слова «перевязь» — «изделие легкой промышленности; изделие; полоса (кусок); кусок (отдельная часть); повязка, повязка на теле;...» при верном ответе «медицинская повязка; медицинская продукция».

4. Слишком конкретные кандидаты. Такие случаи связаны с неполнотой ресурса, а не с тем, как модель выявляет семантические связи. Примером может служить предсказание Qwen3-8B-Base для слова «бойскаут» — «член организации; участник; человек по роли; учащийся; подросток;...» при правильном ответе «мальчик; мужчина, человек мужского пола; скаут; несовершеннолетние дети». Определение этого слова по словарю [Крысин, 2018] — «мальчик

или подросток — член скаутской (см. скаут) организации», из чего можно сделать вывод о релевантности кандидата на первой позиции.

5. Ошибки, связанные с многозначностью. Предсказание гиперонимов для слов, обозначающих несколько понятий одновременно, вне контекста использования также составляет проблему. В частности, в тестовую выборку входит слово «лигатура», имеющее следующие гиперонимы: (1) «хирургическое оборудование», «нить (предмет)», (2) «металл», «добавление (то, что добавлено)» и (3) «графический знак», «сочетание, комбинация». Моделям YandexGPT-5-8B-Lite и Qwen3-8B-Base удалось предсказать только последнее значение: модель серии YandexGPT смогла предсказать непосредственный гипероним «графический знак», а также «знак, обозначение», а модели серии Qwen3 — гипероним только второго порядка («изображение (результат)»).

6. Интерференция другого языка в русский текст. До обработки вывода для извлечения синсетов, среди кандидатов могут попадаться как частичные, так и полные замены слов: высший «законодательный organ» (LLaMA-3.1-8B); «органическое glass», «geometric solid», «combinatorial analysis», «травянистое植物» (Qwen2.5-7B).

7. Неверно предсказана часть речи в названии синсета. Например, модель RuAdapt-Qwen2.5-7B для слова «осиплость» предсказала гиперонимы «измениться, изменение; изменить, сделать иным;...», вероятно, связав их с событийной семантикой корня «осип-» — ‘стать хриплым’. При этом в RuWordNet для данного слова требовалось предсказать признак: «хриплый; глухой, глухо звучащий».

6. Обсуждение результатов

Проведенное исследование позволяет сделать ряд выводов как относительно эффективности больших языковых моделей в задаче пополнения таксономий, так и относительно методологических ограничений, присущих текущему подходу. Анализ полученных результатов выявляет несколько ключевых закономерностей.

6.1. Ограничения применения универсальных мультязычных моделей

Эксперименты на материале английского языка (WordNet) показали, что дообучение современных мультязычных LLM с 7–8 млрд параметров, даже при использовании методологии TaxoLLaMA, не позволяет превзойти качество лучших ранее предложенных методов. Только модели семейства Qwen (Qwen2.5-7B и Qwen3-8B-Base) смогли обойти базовый метод FastText, однако они по-прежнему значительно уступают подходу на основе мета-эмбеддингов (AAEME). Это наблюдение указывает на то, что простой перенос таксономических знаний через дообучение на парах «гипоним–гипероним» имеет естественные пределы.

Данный факт объясняется рядом методологических причин. Во-первых, обучение на изолированных словах без контекста их употребления (как это сделано в оригинальном TaxoLLaMA) не позволяет модели разрешать многозначность. В реальных условиях новое понятие всегда появляется в некотором контексте, и его отсутствие в обучающей и тестовой выборках ставит модель в заведомо более сложные условия, чем те, в которых действует человек-лексикограф. Во-вторых, структура WordNet, где узлы представлены синсетами, часто включающими несколько слов или уточнения в скобках, делает прямую генерацию корректного названия узла крайне затруднительной. Это вынуждает использовать дополнительные методы пост-обработки на основе векторных представлений, которые, будучи несовершенными сами по себе, накапливают погрешность.

6.2. Критическая роль языковой специализации

Наиболее показательные результаты получены на русскоязычных данных. Все модели, прошедшие специализированную адаптацию для русского языка (серия RuAdapt), продемонстрировали устойчивый и значительный прирост качества по сравнению со своими исходными мультязычными версиями. Прирост по метрике MAP составил от 7.8 до 12.3 процентных пункта, что убедительно подтверждает эффективность методик, учитывающих особенности русской морфологии и лексики на этапе токенизации и предобучения.

Особого внимания заслуживает результат модели YandexGPT-5-Lite-8B-pretrain, которая не только превзошла все остальные дообученные модели, но и установила новый рекорд на задаче пополнения RuWordNet (55.42 MAP против предыдущего лучшего результата 47.4). Этот успех, по-видимому, обусловлен сочетанием нескольких факторов: высоким качеством базовой модели, ее изначальной ориентацией на русский язык, а также, вероятно, более репрезентативным корпусом предобучения, включающим тексты, семантически близкие к тематике RuWordNet. Важно отметить, что, хотя нельзя полностью исключить возможность косвенного присутствия данных RuWordNet в обучающем корпусе YandexGPT, разница в приросте качества после применения векторной пост-обработки у этой модели и у моделей RuAdapt сопоставима, что косвенно свидетельствует об отсутствии прямой утечки данных.

6.3. Сравнение дообучения и контекстного обучения (few-shot)

Эксперименты с коммерческими моделями в режимах zero-shot и few-shot выявили важную закономерность: даже самые современные LLM с большим числом параметров (70B+), но без целевого дообучения, не способны достичь качества специализированных моделей меньшего размера. В режиме zero-shot лучший результат продемонстрировала YandexGPT-5 (36.38 MAP), что, хотя и сопоставимо с некоторыми дообученными мультязычными моделями, значительно уступает лидерам.

Предъявление 10 примеров (few-shot) оказало неоднозначное влияние. С одной стороны, это повысило метрику MRR для YandexGPT-5, что говорит о способности модели лучше идентифицировать первый релевантный гипероним. С другой стороны, метрика MAP для всех моделей снизилась. Это объясняется тем, что примеры, по-видимому, сужают пространство генерации: модель начинает подражать структуре и длине ответов из демонстраций, что приводит к меньшему числу генерируемых кандидатов и, как следствие, снижению полноты. Кроме того, модель не имеет возможности «узнать» из нескольких примеров все разнообразие возможных структур синсетов (включая случаи, где гипероним выражен глаголом или

прилагательным). Таким образом, в условиях ограниченного числа демонстраций few-shot learning уступает полноценному дообучению.

6.4. Анализ ошибок и перспективы улучшения метода

Классификация ошибок, допущенных моделями, позволила не только диагностировать их слабые места, но и наметить пути совершенствования подхода. Ошибки были разделены на четыре основных типа:

Неверное семантическое поле (категориальные ошибки). Характерны для более слабых моделей (например, Gemma-7b) и связаны с непониманием предметной области целевого слова.

Подмена таксономических отношений другими (например, часть-целое, ассоциация). Модели часто путают «быть разновидностью» с «обладать свойством» или «являться частью».

Неверный уровень абстракции. Модели склонны предсказывать либо слишком общие (часто), либо слишком конкретные (реже) гиперонимы. Во втором случае предсказание часто семантически верно, но не соответствует более абстрактному узлу, заложенному в таксономии, что указывает на расхождения между «здравым смыслом» модели и экспертной разметкой WordNet/RuWordNet.

Неполнота при многозначности. Наиболее сложный тип ошибок. Модели, обученные на парах слово → гипероним, как правило, «выучивают» только одно, наиболее частотное значение слова, игнорируя остальные.

Для русскоязычных моделей были выявлены два дополнительных класса проблем, связанных со спецификой данных: интерференция других языков (вкрапления английских, китайских и др. слов в генерацию, особенно у моделей без дообучения) и несоответствие частеречной принадлежности (когда модель генерирует существительные, а целевой синсет в RuWordNet выражен глаголом или прилагательным).

На основе анализа ошибок можно предложить несколько направлений для дальнейшего развития методологии:

Обогащение входных данных. Наиболее перспективным представляется переход от генерации по одному слову к генерации с опорой на контекст употребления. Как показано в [Nikishina et al., 2020a], использование контекстов из корпуса новостей позволяет частично снять проблему многозначности и приблизить задачу к реальному сценарию пополнения таксономий. В этом смысле методология ТахоLLaMA, использующая определения, является шагом в верном направлении, но определения сами по себе часто содержат гипероним, что превращает задачу в упрощенную.

Смена роли LLM: с генератора на ранкер. Ряд исследований [Ma et al., 2023] показывает, что способность LLM к ранжированию и оценке релевантности может быть сильнее, чем способность к генерации в условиях ограниченной информации. Перспективным выглядит подход, при котором первичный список кандидатов генерируется быстрым методом (например, FastText или мета-эмбедингами), а LLM используется для их переранжирования с учетом более сложных семантических и таксономических закономерностей.

Гибридный подход RAG (Retrieval-Augmented Generation). Логичным развитием предыдущей идеи является применение архитектуры, дополненной поиском. Модель получает на вход не только гипоним, но и извлеченные из таксономии или внешних источников контексты, списки потенциальных гиперонимов или определения. Это позволяет модели принимать более обоснованные решения, не полагаясь исключительно на «заученные» из обучающей выборки паттерны.

Таким образом, результаты данной работы подтверждают потенциал больших языковых моделей для автоматического пополнения таксономий, особенно при условии их целевой адаптации к языку и задаче. В то же время, они демонстрируют, что для достижения качества, достаточного для практического применения, необходим переход к более сложным, контекстно-зависимым архитектурам, выходящим за рамки простого дообучения на парах слово-гипероним.

Заключение

В работе впервые исследована применимость генеративных больших языковых моделей к задаче автоматического пополнения таксономий на русском языке. С использованием диахронических версий WordNet и RuWordNet и методологии TaxoLLaMA проведено сравнение мультязычных, специализированных русскоязычных и закрытых коммерческих моделей.

Эксперименты показали, что на английских данных мультязычные LLM с 7–8 млрд параметров уступают лучшим методам на основе мета-эмбедингов. На русскоязычном материале выявлено критическое преимущество моделей, адаптированных под русский язык: все версии, прошедшие адаптацию по методике RuAdapt, значительно превосходили свои исходные мультязычные аналоги.

Наилучший результат достигнут при дообучении YandexGPT-5-Lite-8B-pretrain — 55.42 MAP, что существенно выше предыдущего лучшего показателя 47.4 MAP. Модели, адаптированные по методике RuAdapt, также уверенно превосходили базовый метод FastText. Тестирование закрытых моделей в режимах zero-shot и few-shot подтвердило необходимость целевого дообучения: даже модели с 70B+ параметров без адаптации уступают специализированным моделям меньшего размера.

Анализ ошибок выявил основные проблемы моделей: неверное семантическое поле, подмена отношений, неверный уровень абстракции, неполнота при многозначности, а также специфические для русского языка частеречные несоответствия и языковую интерференцию.

Перспективы дальнейших исследований связаны с переходом от генерации по изолированному слову к использованию контекстов употребления, применению LLM в роли ранкеров и разработке гибридных архитектур RAG, объединяющих поиск информации с генерацией ответа.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Исследование выполнено при финансовой поддержке Междисциплинарной научно-педагогической школы Московского университета (грант № 23-Щ05-11) и государственного задания (регистрационный № 124020100068-4), а также Некоммерческого фонда развития науки и образования «Интеллект».

КОНФЛИКТ ИНТЕРЕСОВ

Авторы данной работы заявляют, что у них нет конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

В данной работе отсутствуют исследования человека или животных.

Список литературы

- Большой толковый словарь русского языка / Гл. ред. С. А. Кузнецов. СПб., 2003.
- Крысин Л.П. Современный словарь иностранных слов. Москва, АСТ-Пресс школа, 2018.
- Лукашевич Н. В. Тезаурусы в задачах информационного поиска. Монография. – 2011.
- Aldine A. I. A. et al. Redefining Hearst Patterns by using Dependency Relations //KEOD. – 2018. – С. 146-153.
- Aly R. et al. Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings //arXiv preprint arXiv:1906.02002. – 2019.
- Anke L. E. et al. Supervised distributional hypernym discovery via domain adaptation //Proceedings of the 2016 conference on empirical methods in natural language processing. – 2016. – С. 424-435.
- Arefyev N. V. et al. Word2vec not dead: predicting hypernyms of co-hyponyms is better than reading definitions //Computational Linguistics and Intellectual Technologies. – 2020. – С. 13-32.

Babaei Giglou H., D'Souza J., Auer S. LLMs4OL: Large language models for ontology learning //International Semantic Web Conference. – Cham : Springer Nature Switzerland, 2023. – C. 408-427.

Bai Y. et al. Hypernym discovery via a recurrent mapping model //Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. – 2021. – C. 2912-2921.

Baroni M., Lenci A. How we BLESSed distributional semantic evaluation //Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. – 2011. – C. 1- 10.

Baroni M. et al. Entailment above the word level in distributional semantics //Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. – 2012. – C. 23-32.

Bernier-Colborne G., Barriere C. Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery //Proceedings of the 12th international workshop on semantic evaluation. – 2018. – C. 725-731.

Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. – " O'Reilly Media, Inc.", 2009.

Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology //Nucleic acids research. – 2004. – T. 32. – №. suppl_1. – C. D267-D270.

Bojanowski P. et al. Enriching word vectors with subword information //Transactions of the association for computational linguistics. – 2017. – T. 5. – C. 135-146.

Bollacker K. et al. Freebase: a collaboratively created graph database for structuring human knowledge //Proceedings of the 2008 ACM SIGMOD international conference on Management of data. – 2008. – C. 1247-1250.

Bollegala D., Bao C. Learning word meta-embeddings by autoencoding //Proceedings of the 27th international conference on computational linguistics. – 2018. – C. 1650-1661.

Bordea G., Lefever E., Buitelaar P. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2) //Proceedings of the 10th international workshop on semantic evaluation (semeval-2016). – 2016. – C. 1081-1091.

Bordes A. et al. Learning structured embeddings of knowledge bases //Proceedings of the AAAI conference on artificial intelligence. – 2011. – Т. 25. – №. 1. – С. 301-306.

Bouraoui Z., Camacho-Collados J., Schockaert S. Inducing relational knowledge from BERT //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 05. – С. 7456-7463.

Brown T. et al. Language models are few-shot learners //Advances in neural information processing systems. – 2020. – Т. 33. – С. 1877-1901.

Cai X. et al. Improving word embeddings by emphasizing co-hyponyms //Web Information Systems and Applications: 15th International Conference, WISA 2018, Taiyuan, China, September 14–15, 2018, Proceedings 15. – Springer International Publishing, 2018. – С. 215-227.

Camacho-Collados J. et al. SemEval-2018 task 9: Hypernym discovery //Proceedings of the 12th international workshop on semantic evaluation. – 2018. – С. 712-724.

Chen C., Lin K., Klein D. Constructing taxonomies from pretrained language models //arXiv preprint arXiv:2010.12813. – 2020.

Cherti M. et al. Reproducible scaling laws for contrastive language-image learning //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2023. – С. 2818-2829.

Chung H. W. et al. Scaling Instruction-Finetuned Language Models //arXiv e-prints. – 2022. – С. arXiv: 2210.11416.

Dementieva D. et al. Methods for detoxification of texts for the russian language //Multimodal Technologies and Interaction. – 2021. – Т. 5. – №. 9. – С. 54.

Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

Espinosa-Anke L. et al. SAVANA: a global information extraction and terminology expansion framework in the medical domain. – 2016.

Farrugia J. P. Exploring hypernym discovery: a projection learning perspective : дис. – University of Malta, 2019.

Fellbaum C. (ed.). WordNet: An electronic lexical database. – MIT press, 1998.

Fu R. et al. Learning semantic hierarchies via word embeddings //Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2014. – C. 1199-1209.

Grover A., Leskovec J. node2vec: Scalable feature learning for networks //Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. – 2016. – C. 855-864.

Gupta S. et al. Deep learning with limited numerical precision //International conference on machine learning. – PMLR, 2015. – C. 1737-1746.

Han L. et al. UMBC_EBIQUITY-CORE: Semantic textual similarity systems //Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. – 2013. – C. 44-52.

Hanna M., Mareček D. Analyzing BERT's knowledge of hypernymy via prompting //Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. – 2021. – C. 275-282.

Hearst M. A. Automatic acquisition of hyponyms from large text corpora //COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics. – 1992.

Hochreiter S., Schmidhuber J. Long short-term memory //Neural computation. – 1997. – T. 9. – №. 8. – C. 1735-1780.

Hu E. J. et al. Lora: Low-rank adaptation of large language models //ICLR. – 2022. – T. 1. – №. 2. – C. 3.

Iacobacci I., Pilehvar M. T., Navigli R. Senseembed: Learning sense embeddings for word and relational similarity //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). – 2015. – C. 95-105.

Jain D., Anke L. E. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction //arXiv preprint arXiv:2202.04876. – 2022.

Jenatton R. et al. A latent factor model for highly multi-relational data //Advances in neural information processing systems. – 2012. – T. 25.

Jiang M. et al. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations //Proceedings of the ACM Web Conference 2022. – 2022. – C. 925-934.

Jiang A. Q. et al. Mistral 7b // arXiv preprint arXiv:2310.06825. –2023.

Kingma D. P. et al. Auto-encoding variational bayes // arXiv:1312.6114v10 [stat.ML]. – 2014.

Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language //arXiv preprint arXiv:1905.07213. – 2019.

Kutuzov A., Kuzmenko E. WebVectors: a toolkit for building web interfaces for vector semantic models //International conference on analysis of images, social networks and texts. – Cham : Springer International Publishing, 2016. – C. 155-161.

Lazaridou A. et al. Mind the gap: Assessing temporal generalization in neural language models //Advances in Neural Information Processing Systems. – 2021. – T. 34. – C. 29348-29363.

Le Scao T. et al. Bloom: A 176b-parameter open-access multilingual language model. – 2023.

Levy O., Goldberg Y., Dagan I. Improving distributional similarity with lessons learned from word embeddings //Transactions of the association for computational linguistics. – 2015. – T. 3. – C. 211-225.

Liao J., Chen X., Du L. Concept understanding in large language models: An empirical study. – 2023.

Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.

Loshchilov I., Hutter F. Decoupled weight decay regularization //arXiv preprint arXiv:1711.05101. – 2017.

Loukachevitch N. V., Dobrov B. V. Development and Use of Thesaurus of Russian Language RuThes //Co-operating Organisations. – 2002. – C. 65.

Loukachevitch N. V. et al. Creating Russian wordnet by conversion //Computational Linguistics and Intellectual Technologies. – 2016. – C. 405-415.

Loukachevitch N. V. Automatic Methods for Extracting Taxonomic Relationships from Texts //Pattern Recognition and Image Analysis. – 2023. – T. 33. – №. 3. – C. 398-406.

Ma Y. et al. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! //arXiv preprint arXiv:2303.08559. – 2023.

Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.

Miller G. A. et al. Introduction to WordNet: An on-line lexical database //International journal of lexicography. – 1990. – T. 3. – №. 4. – C. 235-244.

Moskvoretskii V., Panchenko A., Nikishina I. Are large language models good at lexical semantics? a case of taxonomy learning //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – 2024. – C. 1498-1510.

Moskvoretskii V. et al. TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks //arXiv preprint arXiv:2403.09207. – 2024.

Moskvoretskii V. et al. Large Language Models for Creation, Enrichment and Evaluation of Taxonomic Graphs // Semantic Web Journal (forthcoming). – 2024.

Nakashole N., Weikum G., Suchanek F. PATTY: A taxonomy of relational patterns with semantic types //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – 2012. – C. 1135-1145.

Navigli R., Ponzetto S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network //Artificial intelligence. – 2012. – T. 193. – C. 217-250.

Nickel M., Kiela D. Poincaré embeddings for learning hierarchical representations //Advances in neural information processing systems. – 2017. – T. 30.

Nikishina I. et al. RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian language //arXiv preprint arXiv:2005.11176. – 2020

Nikishina I. et al. Studying taxonomy enrichment on diachronic wordnet versions //arXiv preprint arXiv:2011.11536. – 2020..

Nikishina I. et al. Evaluation of taxonomy enrichment on diachronic wordnet versions //Proceedings of the 11th Global Wordnet Conference. – 2021. – C. 126-136.

Nikishina I. et al. Taxonomy enrichment with text and graph vector representations //Semantic Web. – 2022. – T. 13. – No. 3. – C. 441-475.

Nikishina I. et al. Cross-modal contextualized hidden state projection method for expanding of taxonomic graphs //Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing. – 2022. – C. 11-24.

Nikishina I. et al. Predicting Terms in IS-A Relations with Pre-trained Transformers //Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings). – 2023. – C. 134-148.

Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – C. 1532-1543.

Petroni F. et al. Language models as knowledge bases? //arXiv preprint arXiv:1909.01066. – 2019.

Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – T. 1. – No. 8. – C. 9.

Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning. – PmLR, 2021. – C. 8748-8763.

Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //Journal of machine learning research. – 2020. – T. 21. – №. 140. – C. 1-67.

Ravichander A. et al. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT //Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics. – 2020. – C. 88-102.

Roller S., Erk K. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment //arXiv preprint arXiv:1605.05433. – 2016.

Roller S., Kiela D., Nickel M. Hearst patterns revisited: Automatic hypernym detection from large text corpora //arXiv preprint arXiv:1806.03191. – 2018.

Sabirova K., Lukanin A. Automatic Extraction of Hypernyms and Hyponyms from Russian Texts //AIST (supplement). – 2014. – C. 35-40.

Santus E. et al. Evaluation 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models //Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications. – 2015. – C. 64-69.

Shen J. et al. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network //Proceedings of the web conference 2020. – 2020. – C. 486-497.

Shwartz V., Santus E., Schlechtweg D. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection //arXiv preprint arXiv:1612.04460. – 2016.

Singh M., Silakari O. Chapter 4—Flavone: an important scaffold for medicinal chemistry //Key heterocycle cores for designing multitargeting molecules. Elsevier, Amsterdam. – 2018.

Sinha A. et al. An overview of microsoft academic service (mas) and applications //Proceedings of the 24th international conference on world wide web. – 2015. – C. 243-246.

Snow R., Jurafsky D., Ng A. Learning syntactic patterns for automatic hypernym discovery //Advances in neural information processing systems. – 2004. – T. 17.

Socher R. et al. Reasoning with neural tensor networks for knowledge base completion //Advances in neural information processing systems. – 2013. – T. 26.

Subramaniam L. V., Nanavati A. A., Mukherjea S. Enriching one taxonomy using another //IEEE transactions on knowledge and data engineering. – 2009. – T. 22. – №. 10. – C. 1415-1427.

Tanev H., Rotondi A. Deftor at SemEval-2016 task 14: Taxonomy enrichment using definition vectors //Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). – 2016. – C. 1342-1345.

Team Google. Gemma: Open models based on gemini research and technology //arXiv preprint arXiv:2403.08295. – 2024.

Tikhomirov M., Chernyshev D. Impact of tokenization on LLaMa Russian adaptation //2023 Ivannikov Ispras Open Conference (ISPRAS). – IEEE, 2023. – C. 163-168.

Tikhomirov M., Chernyshev D. Facilitating large language model Russian adaptation with Learned Embedding Propagation //arXiv preprint arXiv:2412.21140. – 2024.

Tikhomirov M. M., Chernyshev D. I. Improving Large Language Model Russian adaptation with preliminary vocabulary optimization //Lobachevskii Journal of Mathematics. – 2024. – T. 45. – №. 7. – C. 3211-3219.

Tikhomirov M. M., Loukachevitch N. V., Parkhomenko E. A. Combined approach to hypernym detection for thesaurus enrichment //Computational Linguistics and Intellectual Technologies. – 2020. – C. 736-746.

Tikhomirov M., Loukachevitch N. Meta-embeddings in taxonomy enrichment task //Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue. – 2021.

Tikhomirov M., Loukachevitch N. Exploring Prompt-Based Methods for Zero-Shot Hypernym Prediction with Large Language Models //arXiv preprint arXiv:2401.04515. – 2024.

Touvron H. et al. Llama: Open and efficient foundation language models //arXiv preprint arXiv:2302.13971. – 2023.

Trofimova M. V., Arkhipov M. U. HyBert Team at Dialogue Evaluation 2020: Transformer for Hypernym Extraction //Computational Linguistics and Intellectual Technologies. – 2020. – C. 1170-1179.

Ustalov D. et al. Negative sampling improves hypernymy extraction based on projection learning //arXiv preprint arXiv:1707.03903. – 2017.

Vulić I. et al. Hyperlex: A large-scale evaluation of graded lexical entailment //Computational Linguistics. – 2017. – T. 43. – №. 4. – C. 781-835.

de Vroe B., Cornelis S. G. Temporality and modality in entailment graph induction. – 2023.

Weeds J. et al. Learning to distinguish hypernyms and co-hyponyms //Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. – 2014. – C. 2249-2259.

Xu Y. et al. Hyperminer: Topic taxonomy mining with hyperbolic embedding //Advances in Neural Information Processing Systems. – 2022. – T. 35. – C. 31557-31570.

Yamane J. et al. Distributional hypernym generation by jointly learning clusters and projections //Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. – 2016. – C. 1871-1879.

Yang C. et al. Network representation learning with rich text information //IJCAI. – 2015. – T. 2015. – C. 2111-2117.

Yang A. et al. Qwen2. 5 technical report //arXiv preprint arXiv:2412.15115. – 2024.

Yun G. et al. Hypert: hypernymy-aware BERT with Hearst pattern exploitation for hypernym discovery //Journal of Big Data. – 2023. – T. 10. – №. 1. – C. 141.

Zhang J. et al. Graph-bert: Only attention is needed for learning graph representations //arXiv preprint arXiv:2001.05140. – 2020.

Zhang J. et al. Taxonomy completion via triplet matching network //Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – T. 35. – №. 5. – C. 4662-4670.

Zhang S. et al. Opt: Open pre-trained transformer language models //arXiv preprint arXiv:2205.01068. – 2022.