ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 007:681.512.2

ЭМБЕДДИНГИ ИЕРАРХИЧЕСКИХ ЧИСЕЛ ДЛЯ ОБОГАЩЕНИЯ ТРАНСФОРМЕРНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ ВНЕШНИМИ ОНТОЛОГИЧЕСКИМИ ЗНАНИЯМИ

И. Ю. Каширин, д.т.н., профессор кафедры ВПМ РГРТУ Рязань, Россия;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

Рассматривается новый метод векторизации входных естественно-языковых предложений в языковых генеративных LLM моделях. Основой нового метода является алгебра иерархических чисел, используемая в алгоритмах вычисления семантической близости слов и предложений. Метод пригоден для локальных предметных областей и был апробирован в предметной области «политические новости». Для этой локальной области разработаны OWL онтология и соответствующее графическое представление в форме семантической сети с разметкой базовых концептов иерархическими числами. Семантическая сеть включает общий и прикладной уровни. Общая онтология использует ICF+ отношения, дающие возможность упростить полиморфические операции в моделях знаний.

В программной реализации рассмотренного метода используется технология нейронных генеративных сетей с концентрацией внимания DistilBERT. Обогащение знаний предобученной нейронной сети осуществляется с помощью генерации новых семантических эмбеддингов для слов (концептов) естественно-языковых предложений и их встраивания в новую нейронную сеть перед дообучением в выбранной локальной предметной области.

В качестве обучающих корпусов для получения новой нейросетевой модели mIYu-bert v.2.0. взяты общий корпус из репозитория Hugging Face Datasets и локальный корпус материалов, извлеченных автором настоящей статьи из англоязычных политических статей международных электронных средств массовой информации, в частности RT, cnn, TASS, NYTimes, WSJ.

Экспериментальная часть материала основана на применении программного инструментария языков Python v.3 (Anaconda 3), OWL2EL и пакета программ mIYu-bert v.2.0. Последний инструментарий реализован автором материала.

Выполненная серия экспериментов позволяет квалифицировать новый метод применения иерархических чисел в дообучении моделей LLM для вычисления семантического сходства как основу технологии, не уступающей по эффективности имеющимся на сегодняшний день международным аналогам.

***Целью работы** является презентация нового метода обогащения языковых LLM моделей эмбеддингами иерархических чисел на основе OWL онтологий для локальных предметных областей.*

***Ключевые слова**: эмбеддинги иерархических чисел, нейронные DistilBET-модели, анализ естественного языка, онтологические таксономии, семантическое сходство.*

DOI: 10.21667/1995-4565-2025-93-72-82

Введение

Большие языковые нейросетевые модели (LLM, Large Language Models), например, такие как GPT 4.5 [1], RoBERTa-transformers 4.3.0 [2], Claude 3.2 [3], LLaMA 3.2 [4], Yandex/YaLM-100B [5], стали сейчас явными лидерами в области анализа и синтеза естественно-языковых текстов. Этот уровень технологий искусственного интеллекта дает возможность не только

реализовывать вопросно-ответные, аннотирующие и информационно-аналитические системы, но широко используются датасайнтистами и инженерами по знаниям в разработке новых моделей с широким кругом задач. Рассматриваемые модели используют трансформерную архитектуру [6]. Трансформерные модели, в сравнении с другими, выполняют сравнительно небольшое количество шагов. На каждом шаге применяется механизм самовнимания, который моделирует отношения между всеми словами в предложении независимо от их положения. Трансформер сравнивает каждое слово с каждым другим словом в предложении. Таким образом вычисляется оценка внимания для каждого слова. Оценки внимания определяют, насколько каждое из других слов семантически влияет на формирование следующих в предложении слов. Трансформерные языковые модели качественно превосходят рекуррентные и сверточные модели [7].

В то же время общим недостатком больших моделей можно считать гигантские размеры и вычислительную сложность, требующие для их функционирования больших финансовых затрат (вычислительной стоимости) и большой компьютерной памяти. В этой связи задача оптимизации больших моделей связана с целями обеспечения экономичности и доступности моделей с одновременным сохранением быстродействия и качества. Такая оптимизация даст возможность решать сложные интеллектуальные задачи на широком круге устройств, например в облачных сервисах или мобильных гаджетах.

В силу сказанного целью предлагаемого в статье исследования является разработка метода оптимизации больших трансформерных языковых моделей на основе их обогащения априорными онтологическими знаниями, выраженными в форме эмбеддингов иерархических чисел. В качестве предметной области для естественно-языковых текстов выбрана тема «электронные политические новости».

Эмбеддинги как форма представления знаний

Эмбеддинги (Embeddings) являются основной формой представления знаний в трансформерных языковых моделях. Эмбеддингом называется многомерный вектор чисел, в которые преобразуются такие словарные конструкции (токены) как части слов, слова и фразы. Этот вектор может иметь в LLM тысячи измерений, формируемых аналогично идее семантического пространства Ч. Осгуда [8].

Основным функциональным элементом больших языковых моделей является вычисление семантического сходства [9] словарных конструкций. Конструкции соответствуют обобщенным понятиям − концептам. Концепты должны располагаться в семантическом пространстве по принципу семантической близости: чем более близки понятия по смыслу, тем ближе с точки зрения семантической метрики они должны находиться в этом многомерном пространстве. Например, эмбеддинги для понятий «террорист» и «атака» должны располагаться близко друг к другу, также как эмбеддинги для понятий «военный» и «защитник».

Использование эмбеддингов делает возможным использование семантических отношений и операций. Например, если выполнить над семантическими векторами операцию «боевые действия» + «человек», можно получить эмбеддинг, близкий к понятию «военный». Такая операция предполагает, что в семантическом пространстве можно выделить отношение «военный-гражданский».

Эмбеддинги играют главную роль при формировании весов искусственной нейронной сети, которая в LLM состоит из очень большого количества слоев и связей между ними. Активация нейронов и процесс обучения дают возможность настроить весовые коэффициенты сети, которые и являются ее функциональной сутью. Нейронная сеть является имплицитным (использующим вывод производных сущностей), статистическим и распределенным механизмом.

Большие генеративные языковые модели обучаются на корпусах (обучающих текстах), составляющих пентабайты текста, статей, web-страниц, чатов, программ. Обучение определяется его целями: предсказанием для каждой текущей словарной конструкции другой кон-

струкции, которая по набранной статистике должна последовать за текущей или занять свое место в ее середине. Полученная после обучения нейронная сеть неявно содержит грамматику языка, фактографические сведения о мире, семантические отношения между словарными конструкциями и даже шаблонированные стратегии организации текстовых документов.

Для выделения семантически связанных слов в тексте LLM используют «механизм внимания» (Attention Mechanism). Он заключается в определении семантической близости слов, расположенных, возможно, в разных частях текста. Таким образом, знания в больших моделях используют базовое свойство знаний – активность. Контекстно значимые связи между токенами в естественно-языковом тексте формируются динамически, в процессе работы модели. Это позволяет не только выделять родовидовые, причинно-следственные и меронимические таксономии и многоместные отношения, например «субъект-предикат-объект», но и моделировать понимание человеческих чувств, юмора и иронии, а также генерировать текст от имени LLM.

Существующие методы оптимизации больших языковых моделей

Кратко рассмотрим наиболее эффективные из известных методов упрощения LLM.

Квантование (Quantization). Этот метод использует уменьшение точности чисел, представляющих собой веса активаций нейронной сети. Вместо стандартной 32-битной или 16-битной последовательности используются упрощенные форматы, например 8-битовые, 4-битовые целые числа или даже бинарные последовательности. Метод уменьшает размер LLM в четыре и более раз с сохранением точности Потери оцениваются как 2 – 3 %. Наиболее эффективными считаются технологии квантования с обучением (QAT, Quantization Aware Training) и пост-тренингового квантования (PTQ, Post Training Quantization) с последующей калибровкой.

Дистилляция знаний (Knowledge Distillation). Здесь используется замена большой модели на более простую в результате обучения простой модели большой моделью. Малая модель обучается на готовых предсказаниях (логитах). Примерами таких моделей являются DistilBERT и TinyBERT [10].

Пранинг (Pruning). Метод основан на сокращении избыточных или малозначимых весов или даже нейронов LLM. Затем модель дообучается для восстановления точности. Используются нерегулярный (Unstructured) пранинг как удаление отдельных весов и структурный пранинг (Structured pruning) как удаление групп весов, нейронов или даже слоев.

Модификация трансформеров (TM, Transformer Modifications). Здесь используется модификация существующих трансформеров с целью сведения стандартной, чтобы сделать их более эффективными с точки зрения вычислений и памяти. Классические трансформеры имеют квадратичную сложность вычислений в зависимости от длины текстовой последовательности. Новые подходы, например Performer, Linformer, Longformer, снижают эту сложность до линейной. Примерами являются модели с меньшим количеством слоев, голов внимания и/или меньшей размерностью эмбеддингов: MobileBERT, TinyLlama [11].

Параметрически эффективное дообучение (PEFT, Parameter Efficient Fine Tuning). Это методы, которые дают возможность дообучать большие модели для прикладных задач ограниченного объема, не изменяя всех параметров модели. Здесь добавляются или изменяются некоторые, специфические для прикладной области параметры.

LoRA (Low Rank Adaptation). Этот метод является разновидностью PEFT. Он использует встраивание небольших, обучаемых низкоранговых матриц в существующие весовые матрицы трансформера. Дообучаются только эти небольшие матрицы, а остальные веса остаются прежними.

Спекулятивное декодирование (Speculative Decoding). Такое декодирование является технологией, ускоряющей генерацию текста. При этом используется ограниченная, быстрая временная (draft) модель, которая генерирует несколько токенов вперед, после чего более точная модель тестирует и производит валидацию этих токенов не по одному а полным паке-

том. При правильной работе временной модели большая модель одобряет сразу несколько токенов, в противном случае корректирует их и генерация повторяется. Метод в ряде случаев ускоряет вывод в 2 – 7 раз практически без потери качества генерации.

Новый метод эмбеддингов иерархических чисел

Предварительные исследования [9] показали возможность получать эффективные результаты упрощения или обогащения предобученных моделей при использовании готовых онтологических схем для специфической разметки и последующей векторизации естественно-языковых текстов. Существенная модернизация этого подхода заключается в добавлении к существующим в предобученной модели эмбеддингам индексов из иерархических чисел, идентифицирующих место векторизуемого концепта в родовидовой, причинно-следственной и меронимической онтологиях.

Иерархические числа, впервые описанные в [12], представляют собой числа вида [s] a_0 . a_1 . a_2 . … . a_i . … .a_n , где a_i – целые положительные числа из множества {0, 1, 2, 3…}, a s – символ знака «+» или «-», где положительный знак может быть опущен. Например, иерархическими числами являются: 0.1.14.21.0 или -12.3.0.200.4. Эти числа могут быть десятичными, бинарными или использующими другие базовые основания систем счисления. Для оперирования иерархическими числами используются операции алгебры иерархических чисел [13]. Одним из полезных свойств этих чисел является их эффективность при кодировании концептов в таксономиях семантических отношений. Кроме того, операции соответствующей алгебры позволяют быстро вычислять общих предков двух и более концептов, а также генерировать дочерние концепты таксономий и синтезировать множественное наследование в прикладных полиморфических онтологиях для представления знаний [14].

Наиболее эффективными для применения в векторизации текстов при достаточных вычислительных ресурсах являются бинарные иерархические числа [15]. Чтобы использовать технологию, основанную на методе эмбеддингов иерархических чисел, требуется предварительная подготовка данных, заключающаяся в формировании онтологической модели предметной области.

В качестве примера можно рассмотреть предметную область политических новостей с концептуальной локализацией «Атака». Графическое представление в форме семантической сети, синтезированное генеративной моделью LLM GPT 4.0, представлено на рисунке 1.

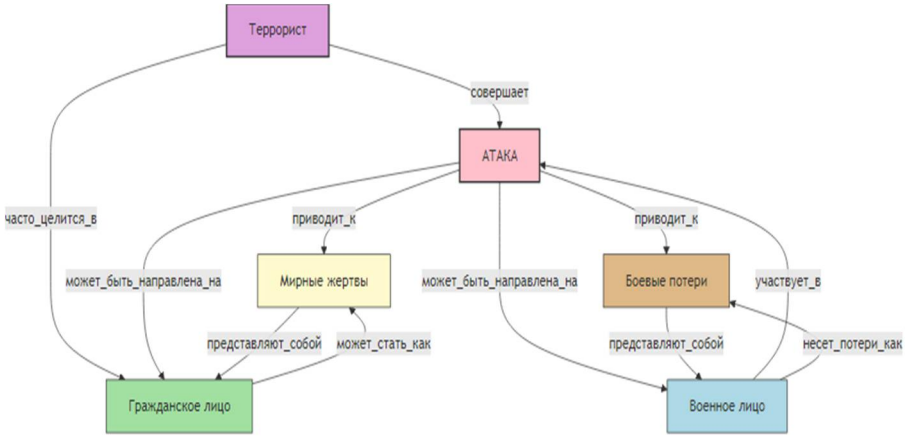

Рисунок 1 – Семантическая сеть в LLM
Figure 1 – Semantic network in LLM

Важно отметить, что в семантической сети выделены концепты онтологии общего уровня, относящиеся к ключевым признакам политических статей выбранной темы: «Атака», «Мирные жертвы», «Боевые потери», «Террорист», «Гражданское лицо», «Мирное лицо». Дуги сети помечены семантическими отношениями, существующими между концептами, например «Атака» «**приводит к**» «Мирные жертвы».

При векторизации текстов на естественном языке определяющую роль играет вычисление семантического сходства слов, понятий, фраз и даже полных текстов. Используя собственные эмбеддинги, LLM вычисляет попарные (каждый с каждым) индексы выделенных в семантической сети концептов, что позволяет сформировать соответствующую тепловую карту (рисунок 2), использующую насыщенность цветового фона для отображения величины смыслового сходства.

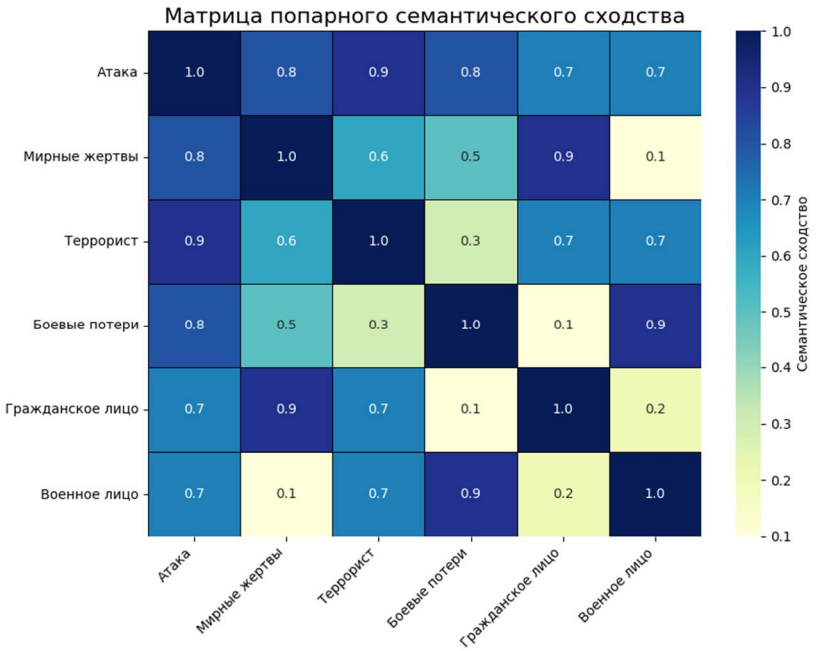

Рисунок 2 – Тепловая карта семантического сходства выделенных понятий, выданная GPT
Figure 2 – Heat map of semantic similarity of selected concepts issued by GPT

Метод эмбеддингов иерархических чисел основан на предварительном создании прикладной онтологии для локальной предметной области. Семантическая сеть, соответствующая предметной области «Атака», представлена на рисунке 3. Далее приводится фрагмент кода на языке OWL2EL [16] для этой онтологии.

```
# Объявление онтологии
:ConflictOntology rdf:type owl:Ontology ;
        rdfs:comment "Пример онтологии для семантической сети концептов предметной области «Атака»."@ru .
    # Основные сущности
    :Атака rdf:type owl:Class ;
        rdfs:label "Атака"@ru .
    :Террорист rdf:type owl:Class ;
        rdfs:label "Террорист"@ru .
    # Общий класс для "лиц"
    :Лицо rdf:type owl:Class ;
        rdfs:label "Лицо"@ru .
    :ГражданскоеЛицо rdf:type owl:Class ;
        rdfs:label "Гражданское лицо"@ru ;
        rdfs:subClassOf :Лицо .          # Гражданское лицо - подкласс Лица
    :ВоенноеЛицо rdf:type owl:Class ;
        rdfs:label "Военное лицо"@ru ;
        rdfs:subClassOf :Лицо .          # Военное лицо - подкласс Лица
    # Классы, представляющие типы потерь, являются подклассами соответствующих типов лиц
    :МирныеЖертвы rdf:type owl:Class ;
        rdfs:label "Мирные жертвы"@ru ;
        rdfs:subClassOf :ГражданскоеЛицо . # Мирные жертвы - подвид Гражданского лица
    :БоевыеПотери rdf:type owl:Class ;
        rdfs:label "Боевые потери"@ru ;
        rdfs:subClassOf :ВоенноеЛицо .   # Боевые потери - подвид Военного лица
```

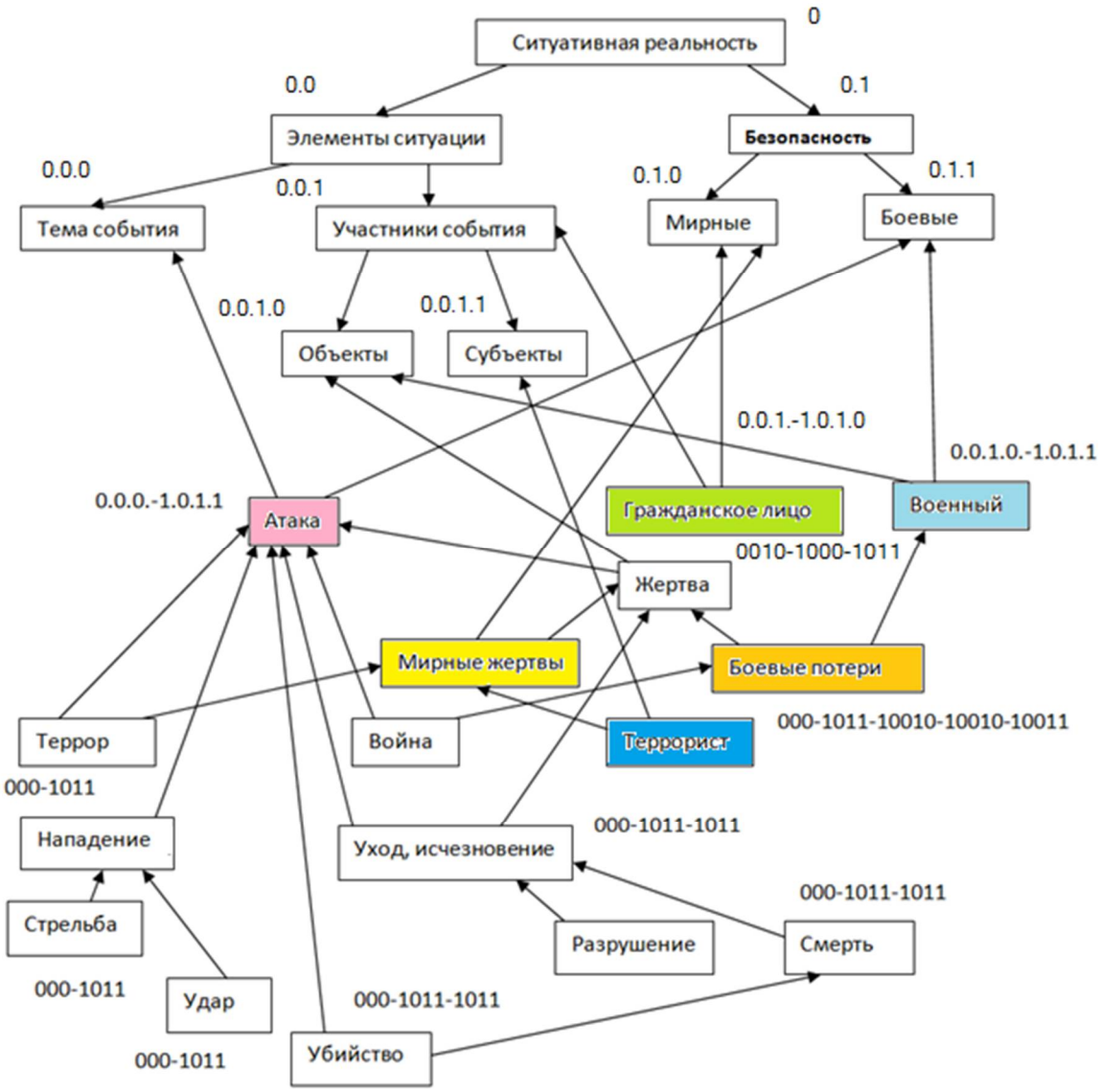

Рисунок 3 – Семантическая сеть с иерархическими числами
Figure 3 – Semantic network with hierarchical numbers

В этом онтологическом описании использованы следующие конструкции:

- owl:Ontology: объявление онтологии для указания ее имени и опционального описания;
- owl:Class: описание концепта в качестве класса в онтологии;
- rdfs:label: хорошо читаемая метка класса для удобства интерпретации;
- rdfs:subClassOf: конструкция формирования таксономии «род-вид» (например, МирныеЖертвы rdfs: subClassOf: ГражданскоеЛицо означает, что «Мирные жертвы» – это подкласс «Гражданского лица»).

Как следует из рисунка 3, отношения между концептами не идентифицируются уникальными наименованиями, но каждый из концептов имеет собственный семантический индекс, представленный иерархическим числом. В верхней части семантической сети (т.е. в верхней подсети) дуги, соединяющие вершины-концепты, соответствуют сложному отношению ICF+[17], которое для всех концептов этой подсети позволяет использовать следующее правило: «любой концепт может быть полиморфически представлен как одна из форм проявления любого другого концепта этой же подсети, но главным отношением, образующим таксономию, является родовидовое отношение IS-A». При этом дуги, выходящие из одной и той же вершины, несут в своей семантике противоположность, которая дает возможность разделить все экземпляры (денотаты) концепта на два непересекающихся подмножества. Таким образом, верхняя семантическая подсеть является общей онтологией. Нижняя подсеть явля-

ется прикладной онтологией, ее дуги направлены вверх и для отличия подсетей иерархические числа ее концептов записаны без разделяющих точек. Как для общей, так и для прикладной онтологии элемент иерархического числа «-1» является разделителем чисел при множественном наследовании и, по сути дела, мог бы быть заменен на запятую.

Для слов естественно-языкового предложения, несущих основной смысл (концентрация внимания), в онтологической сети отыскивается соответствующий концепт. Отличительной чертой прикладной онтологии для локальной предметной области является возможность выбора множества концептов и способа их композиции с концептами общей онтологии. Это реализуется датасайнтистом в соответствии с целями использования результирующей языковой модели с эмбеддингами иерархических чисел. В этом случае появляется возможность рассчитать поправки весовых коэффициентов при вычислении семантической близости слов и предложений естественного языка, используя следующие обозначения.

Пусть есть два слова V и W. Иерархические эбмеддинги этих слов можно представить множествами чисел:

$$V = \{ v_1, v_2, \ldots, v_i, \ldots v_n \}, \quad W = \{ w_1, w_2, \ldots, w_j, \ldots, w_m \}.$$

Можно вычислить пересечение этих множеств $P = V \cap W$. Количество элементов в P (мощность множества) равна |P|.

Определим $L(v_i, w_j)$ как функцию вычисления длины общей части аргументов. Для этого используется операция «°» из алгебры иерархических чисел [18], вычисляющая иерархическое число, соответствующее наиболее общей вершине двух концептов, что интерпретируется как поиск общего предка двух вершин-аргументов, например:

$$v_i = 0.1.1.1,\ w_j = 0.1.0,\ v_i \circ w_j = 0.1.1.1 \circ 0.1.0 = 0.1.0 \circ 0.1.1.1 = 0.1 .$$

Тогда $L(v_i, w_j) = \| 0.1.1.1 \circ 0.1.0 = 0.1.0 \circ 0.1.1.1 \| = \|0.1\| = 2$, где $\| \|$ – операция вычисления длины иерархического числа. Тогда коэффициент семантического сходства двух слов S вычисляется по следующей формуле:

$$S = 2 * \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} 2 * L(v_i, w_j) / (\|v_i\| + \|w_j\|}{|V| + |W|}.$$

Для учета направления поправки коэффициента вводится число μ, которое определяет знак поправочного коэффициента «-» или «+». Например, если общей вершиной таксономии для двух концептов является 0, это означает, что концепты предельно несхожи по своей семантике. Наиболее простым выбором числа μ является длина иерархического числа, соответствующего самой нижней терминальной вершине, деленная пополам. Если $S > \mu$, коэффициент S остается положительным, в противном случае он умножается на -1.

Если необходимо принимать во внимание базовое значение семантического сходства, выдаваемого моделью LLM, поправочный коэффициент, вычисляемый с помощью метода эмбеддингов иерархических чисел, может быть нормализован уменьшением его значения дополнительной поправкой. Например, это может быть умножение на экспериментально подобранное значение.

Экспериментальное исследование

Для получения экспериментальных результатов была разработана программа mIYu-bert 2, обогащающая разработанную ранее естественно-языковую модель mIYu-bert 1 [19] встраиванием в эмбеддинги семантических индексов, представляющих собой иерархические числа для рассмотренной ранее онтологии «политические новости».

Далее представлена схема алгоритма добавки эмбеддингов иерархических чисел.

1) На основе owl-кода и библиотеки word2vec создаётся «словарь семантических индексов» по формату: «входное слово → иерархическое число» (H-id).

2) Создаётся слой нейронной сети mIYu-bert для «эмбеддингов иерархических чисел»: это стандартный nn.Embedding слой, который отображает иерархическое число в плотный вектор. Размерность вектора (semantic_embedding_dim) будет гиперпараметром.

3) Модифицируется вход mIYu-bert:

а) определяются исходные input_ids (токены от токенизатора mIYu-bert);

б) создаются индексы иерархических чисел hierarchic_index_ids, которые соответствуют input_ids по длине;

в) вычисляются стандартные эмбеддинги BERT (word_embeddings + position-tion_embeddings + token_type_embeddings);

г) вычисляются эмбеддинги из уровня semantictic_embedding_layer по hierarchic_index_ids;

д) оба набора эмбеддингов конкатенируются по последней оси (размерности признаков);

е) реализуется проекция объединенного эмбеддинга обратно к hidden_size mIYu-bert с помощью линейного слоя. Это ключевой шаг, так как оригинальные слои mIYu-bert (энкодеры) ожидают вход фиксированной размерности (hidden_size);

ж) новый, обогащенный эмбеддинг подаётся на вход первому слою трансформерного энкодера mIYu-bert.

4) Модель mIYu-bert дообучается на целевой задаче «политические новости». Теперь она обучится использовать семантические иерархические числа и станет новой моделью mIYu-bert 2.

Результатом вычислений в собственной разработке автора статьи mIYu-bert 2 стала тепловая карта, приведенная на рисунке 2. В данном примере был использован дополнительный семантический коэффициент 0,25.

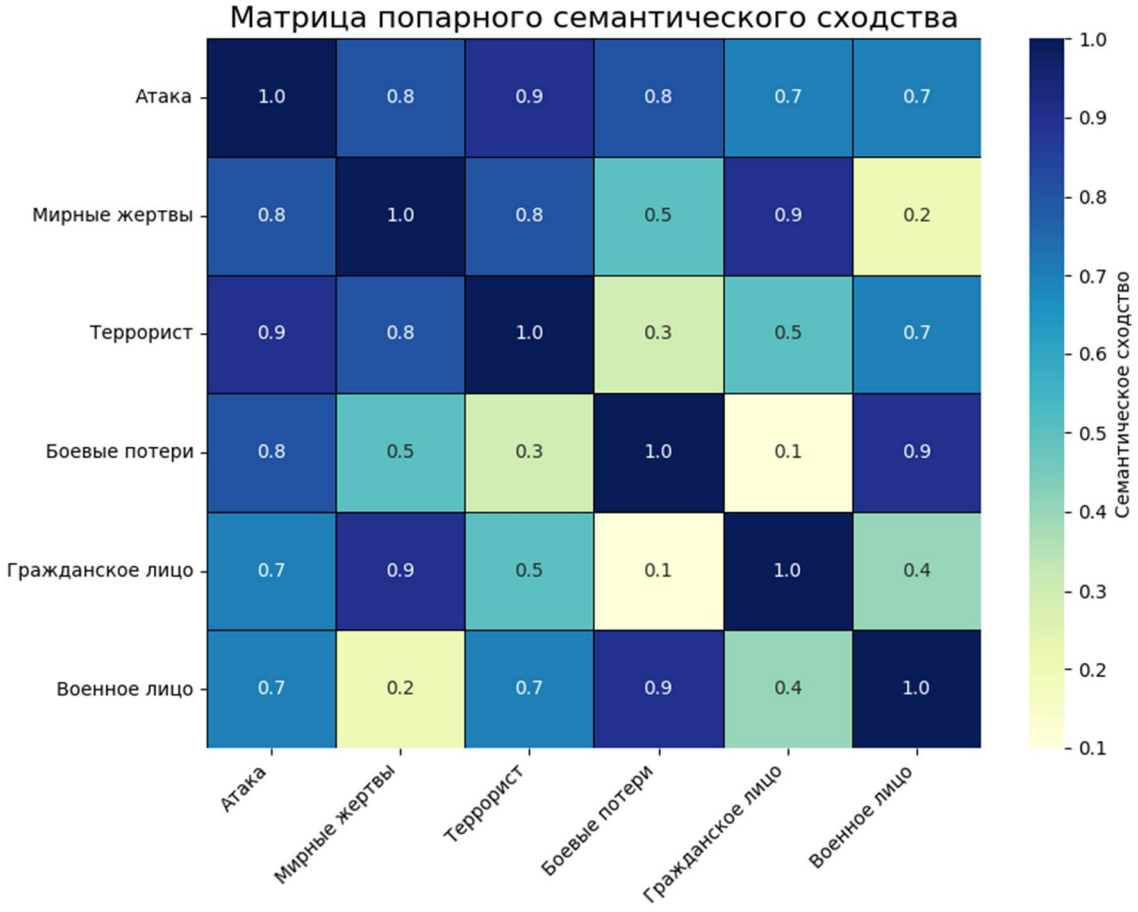

Рисунок 4 – Семантическая сеть с иерархическими числами, выданная mIYu-bert 2
Figure 4 – Heat map of semantic similarity of selected concepts issued by mIYu-bert

Заключение

Произведенные эксперименты позволяют сделать вывод о незначительном изначальном проигрыше модели mIYu-bert 2 при вычислении смыслового сходства одинаковых по смыслу предложений в сравнении с предобученной моделью DistilBERT на корпусе для обучения Hugging Face Datasets [20]. Однако mIYu-bert 2 ощутимо выигрывают при сравнении противоположных или далеких друг от друга по смыслу предложений на подмножестве корпуса из локальной области «политические новости».

В последнем случае модифицированная с использованием иерархических чисел модель mIYu-bert 2 улучшает качество вычисления семантического сходства предложений в сравнении с последней версией языковой модели DistilBERT-base-cased примерно на 3 – 5 %. При этом корректировка результатов осуществляется как в сторону увеличения, так и в сторону уменьшения семантического сходства там, где это оправдано в рамках экспертного мнения.

Библиографический список

1. **Thompson A.D.** GPT-4.5 [Electronic resource]. Update date: January 2024, February 2025.URL: https://lifearchitect.ai/gpt-4-5/ (date of application: 08.05.2025).

2. **Nur Azizah S.F., Cahyono H.D., Sihwi S.W**. Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection. arXiv 2024, arXiv:2202.01907.

3. **Glifton G.** (January 3, 2024). Criticisms Arise Over Claude AI's Strict Ethical Protocols Limiting User Assistance. Light Square. Archived from the original on January 3, 2025. Retrieved January 23, 2024.

4. **Gonçalves J., Silva M., Cabra B.** et al. Evaluating LLaMA 3.2 for Software Vulnerability Detection. [Submitted on 10 Mar 2025]. arXiv:2503.07770. https://doi.org/10.48550/arXiv.2503.07770.

5. Yandex/YaLM-100B. [Electronic resource]. Update date: 30.12.2024, URL: https://github.com/yandex/YaLM-100B. (date of application: 02.06.2025).

6. **Patrick E., Sumith K., Blattmann A.** et al. (2024-03-05), Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, arXiv:2403.03206.

7. **Николенко С., Кадурин А., Архангельская Е.** Глубокое обучение. СПб. Питер, 2018. 480 с.

8. **Osgood Ch.E**. The nature and measurement of meaning // Psychol.Bull. 1952.V. 49.

9. **Каширин И.Ю.** Теория иерархических чисел в задачах вычисления семантического сходства естественно-языковых конструкций. Вестник рязанского государственного радиотехнического университета. 2024. № 88. С. 38-52. DOI: 10.21667/1995-4565-2024-88-38-52.

10. **Hinton G., Vinyals O., Dean J.** 2015 the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

11. **Sun Z., Hongkun Yu., Song X.** et al. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. [Submitted on 6 Apr 2020 (v1), last revised 14 Apr 2020 (this version, v2)]. arXiv:2004.02984v2. https://doi.org/10.48550/arXiv.2004.02984.

12. **Каширин И.Ю.** Иерархические числа для проектирования ICF-таксономий искусственного интеллекта // Вестник Рязанского государственного радиотехнического университета. 2020. № 71. С. 71-82. DOI: 10.21667/1995-4565-2020-71-71-82.

13. **Kashirin I.Yu., Chystyakov P.A.** Intelligent diagnostics using the algebra of hierarchical numbers. IIASU'23 – Artificial intelligence in management, control, and data processing

14. Systems. Proceedings of the II All-Russian scientific conference (Moscow, April 27–28, 2023): In 5 volumes. Moscow, Publishing House «KDU», 2023. Vol. 2. Electronic edition. URL: https://bookonlime.ru/node/72807/ DOI: 10.31453/kdu.ru.978-5-7913-1352-2-2023-406. P-71-75.

15. **Каширин И.Ю.** Применение теории иерархических чисел в проектировании icf-таксономии для оптимизации нейронных сетей // Вестник Рязанского государственного радиотехнического университета. 2022. № 80. С. 118-126. DOI: 10.21667/1995-4565-2022-80-118-126.

16. **Каширин И.Ю.** Двоичные иерархические числа для расчета семантической близости предложений естественного языка // Вестник Рязанского государственного радиотехнического университета. 2023. № 86. С. 110-121.

17. **Mendez J., Suntisrivaraporn B**. Reintroducing CEL as an OWL 2 EL Reasoner. Theoretical Computer Science, TU Dresden, Germany {mendez,meng}@tcs.inf.tu-dresden.de. P.11.

18. **Каширин И.Ю.** Векторизация текста на основе ICF+ онтологии в ансамблях моделей машинного обучения для классификации электронных ресурсов. Вестник Рязанского государственного радиотехнического университета. 2024. № 90. С.41-53. DOI: 10.21667/1995-4565-2024-90-41-53.

19. **Kashirin I.Yu.** Theory of hierarchical numbers in calculation problems semantic similarity of natural language constructions [Electronic resource]. Update date: 30.12.2024, URL: https://kashirin.net/THEORY_OF_HIERARCHICAL_NUMBERS.pdf (date of application: 02.06.2025).

20. **Kashirin I.Yu.** A neural network for classifying media into western and eastern. [Electronic resource]. Update date: 30.12.2024, URL: https://kashirin.net (date of application: 02.06.2025).

21. Hugging Face Datasets. [Electronic resource]. Update date: 20.02.2025, URL: https://huggingface.co/datasets (date of application: 01.04.2025).

UDC 007:681.512.2

EMBEDDINGS OF HIERARCHICAL NUMBERS TO ENRICH TRANSFORMATIONAL LANGUAGE MODELS WITH EXTERNAL ONTOLOGICAL KNOWLEDGE EXTERNAL ONTOLOGICAL KNOWLEDGE

I. Yu. Kashirin, Dr. in technical sciences, full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

A new method for the vectorization of input natural language sentences in language generative LLM models is considered. The basis for a new method is the algebra of hierarchical numbers used in algorithms to calculate semantic proximity of words and sentences. The method is suitable for local subject areas and has been tested in «political news» subject area. OWL ontology and corresponding graphical representation in the form of semantic network with hierarchical numbers marking up basic concepts have been developed for this local area. Semantic network includes general and applied layers. General ontology uses ICF+ relations which make it possible to simplify polymorphic operations in knowledge models.

Software implementation of the method considered uses the technology of neural generative networks with DistilBERT concentration of attention. Knowledge enrichment of pre-trained neural network is carried out by generating new semantic embeddings for words (concepts) of natural language sentences and embedding them into a new neural network before further training in a selected local subject area.

General corpus from Hugging Face Datasets repository and local corpus of materials extracted by the author of this article from English-language political articles of international electronic media, in particular, RT, CNN, TASS, NYTimes, WSJ, are used as training corpora for obtaining a new neural network model mIYu-bert v.2.0.

The experimental part of material is based on Python v.3 (Anaconda 3), OWL2EL, and mIYu-bert v.2.0 software package. The latter toolkit is implemented by the author of the material.

The performed series of experiments allows us to qualify a new method of using hierarchical numbers in further training of LLM models to calculate semantic similarity as the basis for the technology that is not inferior in efficiency to international analogues available today.

***The aim** of this paper is to present a new method for enriching LLM language models with embeddings of hierarchical numbers based on OWL ontologies for local subject areas.*

***Keywords**: embeddings of hierarchical numbers, neural DistilBET models, natural language analysis, ontological taxonomies, semantic similarity.*

DOI: 10.21667/1995-4565-2025-93-72-82

References

1. **Thompson A.D.** GPT-4.5 [*Electronic resource*]. Update date: January 2024, February 2025.URL: https://lifearchitect.ai/gpt-4-5/ (date of application: 08.05.2025).

2. **Nur Azizah S.F., Cahyono H.D., Sihwi S.W.** Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection. *arXiv 2024*, arXiv:2202.01907.

3. **Glifton G.** (January 3, 2024). Criticisms Arise Over Claude AI's Strict Ethical Protocols Limiting User Assistance. *Light Square. Archived from the original on January 3*, 2025. Retrieved Jan-uary 23, 2024.

4. **Gonçalves J., Silva M., Cabra B.** et al. Evaluating LLaMA 3.2 for Software Vulnerability Detection. [Submitted on 10 Mar 2025]. *arXiv:2503.07770*. https://doi.org/10.48550/arXiv.2503.07770.

5. Yandex/YaLM-100B. [*Electronic resource*]. Update date: 30.12.2024, URL: https://github.com/yandex/YaLM-100B. (date of application: 02.06.2025).

6. **Patrick E., Sumith K., Blattmann A.** et al. (2024-03-05), Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, *arXiv*:2403.03206.

7. **Nikolenko S., Kadurin A., Arkhangel'skaya E.** Glubokoe obuchenie. SPb. Piter, 2018. 480 p.

8. **Osgood Ch.E.** The nature and measurement of meaning. *Psychol.Bull.* 1952, vol. 49.

9. **Kashirin I.Yu**. Teoriya ierarkhicheskikh chisel v zadachakh vychisleniya semanticheskogo skhodstva estestvenno-yazykovykh konstrukcij. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2024, no. 88. pp. 38-52. DOI: 10.21667/1995-4565-2024-88-38-52. (in Russian).

10. **Hinton G., Vinyals O., Dean J.** 2015 the knowledge in a neural network. *arXiv preprint arXiv*:1503.02531.

11. **Sun Z., Hongkun Yu., Song X. et al**. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. [Submitted on 6 Apr 2020 (v1), last revised 14 Apr 2020 (this version, v2)]. *arXiv:*2004.02984v2. https://doi.org/10.48550/arXiv.2004.02984.

12. **Kashirin I.Yu**. Ierarkhicheskie chisla dlya proektirovaniya ICF-taksonomij iskusstvennogo intellekta. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2020, no. 71, pp. 71-82. DOI: 10.21667/1995-4565-2020-71-71-82. (in Russian).

13**. Kashirin I.Yu., Chystyakov P.A.** Intelligent diagnostics using the algebra of hierarchical numbers. *IIASU'23 – Artificial intelligence in management, control, and data processing*

14. Systems. *Proceedings of the II All-Russian scientific conference* (Moscow, April 27–28, 2023): In 5 volumes. Moscow, Publishing House «KDU», 2023, vol. 2. Electronic edition. URL: https://bookonlime.ru/node/72807/ DOI: 10.31453/kdu.ru.978-5-7913-1352-2-2023-406. R-71-75.

15. **Kashirin I.Yu**. Primenenie teorii ierarkhicheskikh chisel v proektirovanii icf-taksonomii dlya optimizacii nejronnykh setej. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2022, no. 80, 2022, pp. 118-126. DOI: 10.21667/1995-4565-2022-80-118-126. (in Russian).

16. **Kashirin I.Yu**. Dvoichnye ierarkhicheskie chisla dlya rascheta semanticheskoj blizosti predlozhenij estestvennogo yazyka. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2023, no. 86, pp. 110-121. (in Russian).

17. *Julian Mendez and Boontawee Suntisrivaraporn. Reintroducing CEL as an OWL 2 EL Reasoner*. Theoretical Computer Science, TU Dresden, Germany {mendez,meng}@tcs.inf.tu-dresden.de. P.11.

18. **Kashirin I.Yu**. Vektorizaciya teksta na osnove ICF+ ontologii v ansamblyakh modelej mashinnogo obucheniya dlya klassifikacii ehlektronnykh resursov. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2024, no. 90, pp. 41-53. DOI: 10.21667/1995-4565-2024-90-41-53. (in Russian).

19. **Kashirin I.Yu**. Theory of hierarchical numbers in calculation problems semantic similarity of natural language constructions [*Electronic resource*]. Update date: 30.12.2024, URL: https://kashirin.net/THEORY_OF_HIERARCHICAL_NUMBERS.pdf (date of application: 02.06.2025).

20. **Kashirin I.Yu**. A neural network for classifying media into western and eastern. [*Electronic resource*]. Update date: 30.12.2024, URL: https://kashirin.net (date of application: 02.06.2025).

21. Hugging Face Datasets. [*Electronic resource*]. Update date: 20.02.2025, URL: https://huggingface.co/datasets (date of application: 01.04.2025).