T

Проблемы искусственного интеллекта. 2026. N 1 (40). С. 89-100
Problems of Artificial Intelligence. 2026;1(40):89-100.
Искусственный интеллект и машинное обучение
Научная статья

УДК 004.89
doi: 10.24412/2413-7383-2026-1-40-89-100

Топпер А. М., Гончарова А. Б.
Санкт-Петербургский государственный университет
Университетская наб., д. 7-9, Санкт-Петербург, 199034, Российская Федерация

УЛУЧШЕНИЕ СТРУКТУРИЗАЦИИ ПОНЯТИЙ В ЯЗЫКОВЫХ МОДЕЛЯХ ПОСРЕДСТВОМ ДОМЕННОЙ АДАПТАЦИИ

Topper A. M., Goncharova A. B.
St. Petersburg State University, St Petersburg, Russia
7-9, Universitetskaya Emb., St Petersburg,199034, Russia

IMPROVING THE STRUCTURING OF CONCEPTS IN LANGUAGE MODELS THROUGH DOMAIN ADAPTATION

Актуальность обусловлена необходимостью интерпретации семантических представлений в языковых моделях, особенно в медицинской области. Библиотека Semgeom разработана для анализа геометрии векторных представлений слов методом семантических осей. Проведен сравнительный анализ моделей RuBioRoBERTa и ruRoberta-large, показавший преимущество доменно-адаптированной модели в точности семантической структуризации медицинских понятий. Результаты демонстрируют практическую ценность метода для валидации языковых моделей в специфических профессиональных доменах.
Ключевые слова: языковые модели; семантические оси; доменная адаптация; медицинские эмбеддинги; интерпретируемость ИИ

The relevance of the study is determined by the need to interpret semantic representations in language models, especially in the medical field. The Semgeom library was developed for analyzing embedding geometry using the semantic axes method. A comparative analysis of RuBioRoBERTa and ruRoberta-large models showed the advantage of the domain-adapted model in the accuracy of semantic structuring of medical concepts. The results demonstrate the practical value of the method for validating language models in professional domains.
Keywords: language models; semantic axes; domain adaptation; medical embeddings; AI interpretability

Введение

Модели семейства трансформер (*transformer*) [1] стали стандартом в области обработки естественного языка. Трансформеры создают контекстуальные эмбеддинги слов через многослойную сеть, основанную на механизме самовнимания (self-attention). Каждый токен из исходного предложения на естественном языке последовательно кодируется в векторную форму, при этом учитываются зависимости между всеми словами во входной последовательности. К эмбеддингам добавляются позиционные кодировки для сохранения информации о порядке слов в тексте. Каждый блок трансформера включает механизм многоголового внимания *(multi-head attention*), который позволяет модели одновременно учитывать разные представления входа, и позиционно-зависимый полносвязный слой. Таким образом, выходом модели являются многомерные эмбеддинги слов, в которых зашифрованы как лексические, так и семантические свойства.

В основе этих моделей лежит процесс преобразования текстовых токенов в векторы (эмбеддинги). Эти эмбеддинги формируют семантическое пространство, где геометрическая близость векторов отражает семантическую близость соответствующих слов или выражений. Семантические отношения между понятиями (например, синонимия, антонимия, гипонимия) могут быть выражены через векторные операции.

Актуальность интерпретации внутренних представлений крупных языковых моделей [2-4] особенно высока в медицине, где семантическая адекватность наиболее важна [5]. Поскольку стандартные метрики не отражают предметных знаний модели, ключом к интерпретируемости становится анализ геометрии эмбеддингов [6-8].

Для решения этой проблемы авторами была разработана библиотека Semgeom [9], которая позволяет измерять и визуализировать организацию семантического пространства модели. Библиотека фокусируется на методе семантических осей [10-13] — направлений в пространстве эмбеддингов, соответствующих ключевым бинарным противопоставлениям. Целью является сравнительный анализ того, как доменная адаптация влияет на структуризацию медицинских понятий в семантических пространствах моделей RuBioRoBERTa и ruRoberta-large.

В этой статье проводится сравнительный анализ двух русскоязычных моделей: RuBioRoBERTa [14] (специализированная медицинская модель) и ruRoberta-large [15] (общая доменная модель). Сравнение проводится по семантическим осям, важным в медицине (например, «лечение–симптом» и «диагностика–терапия»). Особое внимание уделяется тому, как доменная адаптация влияет на выраженность семантической структуры в эмбеддингах [16].

Метод

Библиотека Semgeom служит инструментом для анализа семантической структуры эмбеддингов. Для каждой выбранной семантической оси определяется вектор признака: функция feature_direction усредняет разность эмбеддингов слов-полюсов. В словах полюсах одно условно называется «положительным» словом, другое «отрицательным» словом. Полученный вектор нормируется и используется как направление оси. Далее вычисляются проекции эмбеддингов набора слов на этот вектор, дающие численную оценку ассоциации слова с данным семантическим признаком. На практике для каждой оси было задано два множества слов-полюсов, и сформирован вектор признака как разность средних эмбеддингов этих множеств.

Метод семантических осей основан на определении направления признака в векторном пространстве модели. Для бинарного семантического признака задаются два множества слов-полюсов: положительное (P) и отрицательное (N). Каждому слову (w) ставится в соответствие векторное представление слова $v(w) \in R^n$, где n — размерность пространства эмбеддингов модели. Семантическая ось представляет собой нормированный вектор $\vec{d} \in R^n$, вычисляемый на основе этих множеств. При работе с разными моделями полюса для осей и сами исследуемые признаки должны вычисляться только текущей моделью.

Семантическое направление определяется как средняя разность векторов положительных и отрицательных примеров. На основе всех попарных разностных векторов строится множество:

$$D = \left\{ \overrightarrow{v_{p_i}} - \overrightarrow{v_{n_j}}, \quad p_i \in P, n_j \in N \right\}.$$

Для заданных множеств слов-полюсов — положительной размерности m и отрицательной размерности k вектор направления признака вычисляется как:

$$\vec{d} = \frac{1}{m \cdot k} \sum_{i=1}^{m} \sum_{j=1}^{k} \left(\overrightarrow{v_{p_i}} - \overrightarrow{v_{n_j}} \right),$$

где:

$\overrightarrow{v_w}$ – векторное представление слова w,

m, k — количества положительных и отрицательных слов соответственно.

Данная формула усредняет «направления», связывающие каждый положительный пример с каждым отрицательным — это необходимо для выявления общего направления, которое отличает положительные примеры от отрицательных.

Далее направление нормируется:

$$\hat{d} = \frac{\vec{d}}{\|\hat{d}\|}.$$

Таким образом, $\hat{d}$ является единичным вектором, задающим ось признака в семантическом пространстве модели. Нормирование переводит вектор в единичный, чтобы сделать значения проекций сопоставимыми между разными осями и моделями, иначе абсолютная длина вектора зависела бы от размера векторного представления, используемого в исследуемой модели. Данный подход обеспечивает устойчивость к выбросам за счет усреднения по всем возможным парам полюсов.

Масштабируемая проекция для слова w из словаря слов W размерности l:

$$proj_{raw}(w) = \langle \overrightarrow{v(w)}, \hat{d} \rangle.$$

Это классическая ортогональная проекция в линейном пространстве. Благодаря ее использованию, по знаку получившегося значения можно определить к какому полюсу относится исследуемое слово, а по его абсолютной величине — насколько близко слово расположено к полюсу.

Чтобы значения были сопоставимы между осями и моделями, используется линейное масштабирование. Пусть $proj_{raw}(w_i)$ — набор проекций всех терминов W.

Введем следующие обозначения:

$$p_{min} = \min_{w_i \in W} \left(proj_{raw}(w_i) \right), \quad p_{max} = \max_{w_i \in W} \left(proj_{raw}(w_i) \right).$$

Тогда финальная нормированная оценка считается таким образом:

$$proj_{scaled}(w) = \frac{proj_{raw}(w) - p_{min}}{p_{max} - p_{min}} \cdot (r_{max} - r_{min}) + r_{min},$$

где r_{min}, r_{max} — желаемый диапазон, в данном эксперименте они равны соответственно -1 и 1. Такая оценка называется минимаксной (Min-Max). Ее использование

позволяет сохранить относительное расположение значений после проекции. Эта оценка не делает распределение «нормальным», что важно для интерпретации векторных моделей, так как их проекции часто негауссовы [17], [18]. Следственно, этот способ позволяет строить чёткие, ограниченные оси, удобные для визуализации.

Таким образом, предложенный метод:

– создаёт инвариантную, интерпретируемую ось признака, основанную на наборе доменных терминов;
– позволяет представлять доменные различия как проекции на единичный вектор, отражающие «семантическое направление»;
– нормирует значения Min–Max для сопоставимости между моделями;
– сохраняет порядок и интерпретацию расстояний;
– к масштабированию векторов разных моделей.

После того как для каждого слова вычислены проекции на направление признака анализируется структура этих проекций с помощью метрик интерпретируемости, которые отражают как силу выраженности оси, так и однородность ассоциаций слов. В работе используются две основные метрики: mean_abs — среднее абсолютное значение проекций (показывает, насколько сильно в среднем слова смещаются вдоль оси), и std – стандартное отклонение этих проекций (оценка разброса).

$$mean\_abs = \frac{1}{|W|}\sum_{w \in W}|proj(w)|,$$

$$std = \sqrt{\frac{1}{|W|}\sum_{w \in W}(proj(w) - \widehat{proj})^2}, \widehat{proj} = \frac{1}{|W|}\sum_{w \in W} proj(w).$$

Высокое значение среднего абсолютного значения проекций указывает, что соответствующая ось ярко выражена в данной модели, и слова, относящиеся к положительному и отрицательному полюсам, хорошо разделяются. Значение стандартного отклонения этих проекций показывает, насколько однородны ассоциации слов по этой оси. Малое стандартное отклонение означает, что все слова примерно одинаково ассоциированы с осью, т.е. модель «однозначно» понимает эту семантическую категорию. Большое стандартное отклонение проекций говорит о разнообразии ассоциаций, что может отражать размытость концепта в корпусе.

Стоит отметить, так как проекции нормализуются, то обе метрики зависят только от относительных смещений слов вдоль оси, а не от абсолютной длины вектора признака. В интерпретации результатов высокое среднее абсолютное значение проекций вместе с умеренным стандартным отклонением проекций указывает на четко выраженную ось с согласованной группировкой слов. А низкое среднее абсолютное значение проекций или высокое стандартное отклонение проекций могут сигнализировать о необходимости дополнительного доменного обучения модели.

Для эксперимента используются предобученные модели RuBioRoBERTa и ruRoberta-large. Каждая модель применяется к одинаковому набору слов, выбранных по медицинской тематике и общей лексике. Векторное представление слов извлекаются на уровне слоя встраивания (*Embedding layer*), что соответствует представлению каждого токена в словаре. Анализ фокусируется именно на пространстве эмбеддингов, так как векторные представления токенов напрямую связываются с семантическими признаками и не зависят от конкретного контекста.

При работе с неанглоязычными языковыми моделями важно учитывать возможные систематические искажения, возникающие из-за неоднородности корпусов и различий между доменами. Модели могут по-разному кодировать одни и те же медицинские понятия в зависимости от частоты употребления и контекста в корпусе.

Это особенно заметно для русского языка, где специализированных биомедицинских текстов значительно меньше, чем в английском. Подобные эффекты подробно рассматриваются в обзоре методов выявления предвзятостей в неанглоязычных моделях [19].

Результаты эксперимента

Для анализа семантических представлений медицинских понятий в языковых моделях были выбраны две экспертно определённые семантические оси с соответствующими словами-полюсами:

- **«Острое – Хроническое»**:
  - *Положительные образцы (Острое)*: "инфаркт", "инсульт", "перелом", "острая боль", "туберкулёз";
  - *Отрицательные образцы (Хроническое)*: "диабет", "астма", "артрит", "гипертония", "ожирение", "депрессия".
- **«Эндокринные – Нервные»**:
  - *Положительные образцы (Эндокринные)*: "гипотиреоз", "диабет", "гипертиреоз", "аденома", "ожирение";
  - *Отрицательные образцы (Нервные)*: "мигрень", "инсульт", "эпилепсия", "невроз", "остеохондроз".

Для каждой оси был сформирован целевой словарь из релевантных медицинских терминов. Для оси «Острое-Хроническое» использовались такие термины как: "ангина пекторис", "рак молочной железы", "цирроз печени" и др. Для оси «Эндокринные-Нервные» отбирались заболевания с преимущественно эндокринным ("синдром поликистозных яичников", "болезнь Аддисона") или неврологическим ("энцефалит", "болезнь Паркинсона") происхождением.

Для анализа вычислялись проекции каждого термина из целевого словаря на соответствующие семантические оси в векторных пространствах моделей RuBioRoBERTa и ruRoberta.

Визуализация эмбеддингов предоставляет наглядный способ выявлять макроструктуру концептуального пространства и подтвердить согласованность осей. При этом следует учитывать, что двумерные проекции неизбежно упрощают исходное распределение. Эти аспекты подробно рассматриваются в обзорах по визуализации медицинских концептов [20] и сравнении общих и биомедицинских векторных пространств [21]. В российской литературе также подчёркивается значимость корректной визуальной аналитики и адаптации интерфейсов при работе с разными моделями [22], [23].

В биомедицинских моделях концепты разделяются заметно чётче, так как специализированные корпуса создают плотные группы медицинских терминов. Похожие идеи применяются и в современных системах, учитывающих онтологии [24], которые показывают преимущества при поиске и объединении медицинских понятий. Это согласуется с работами, где показаны систематические различия между биомедицинскими и общими пространствами представления терминов [21].

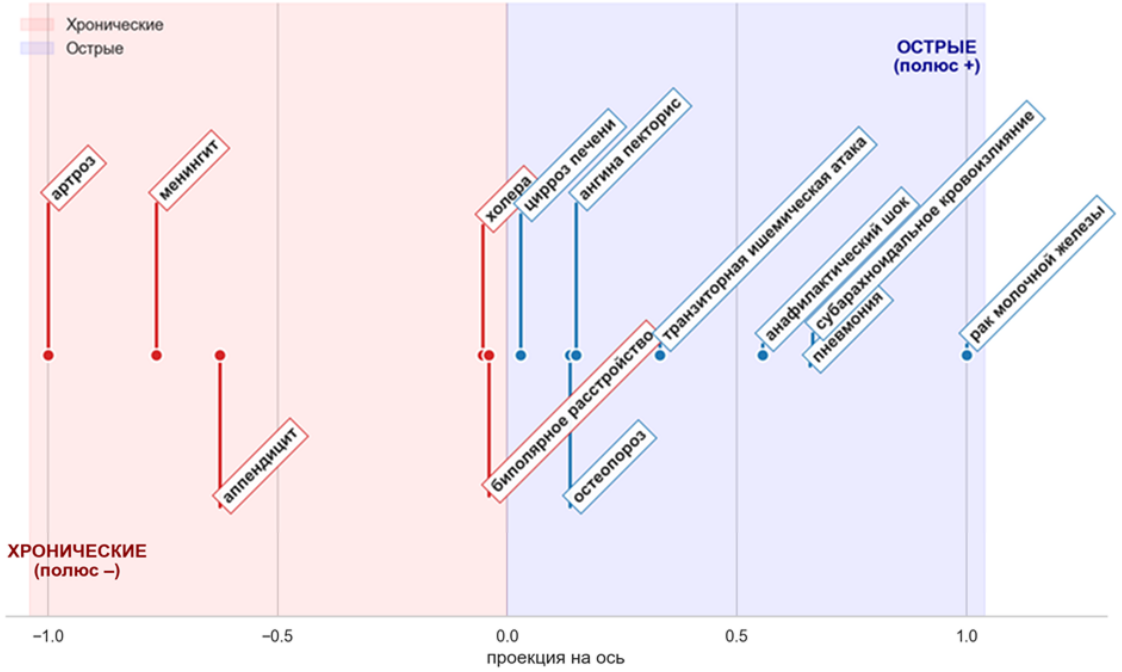

Рисунок 1 – Разделение состояния заболеваний на хронические и острые с использованием модели RuBioRoBERTa

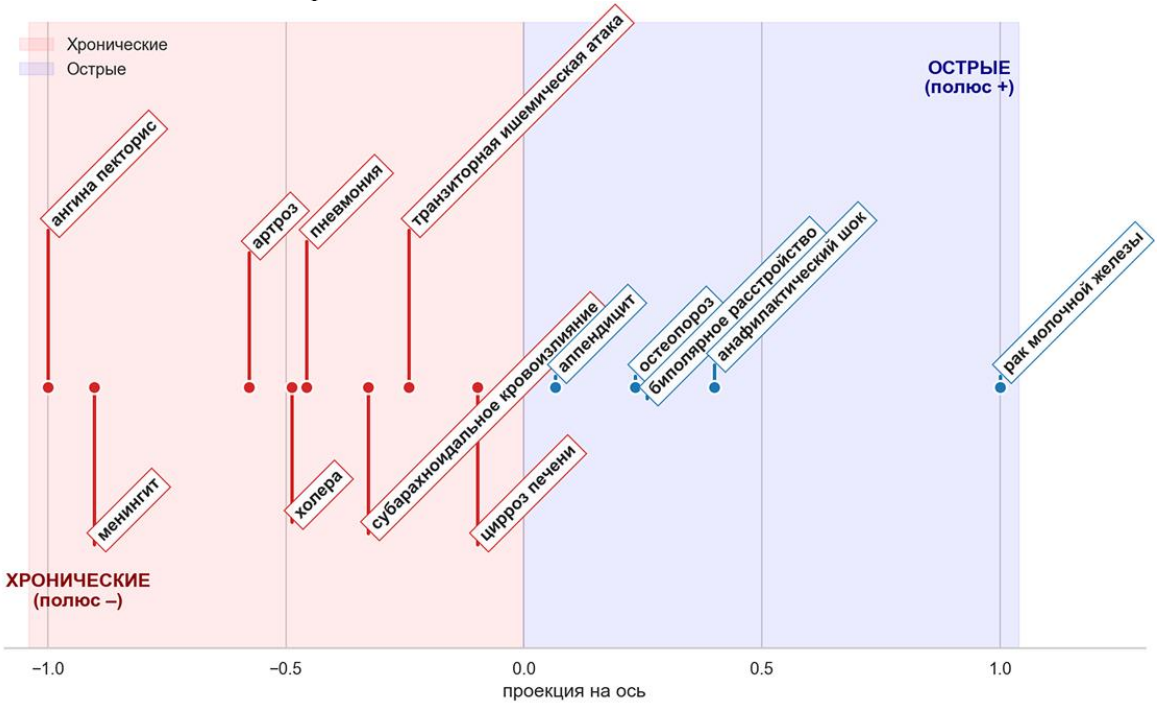

Рисунок 2 – Разделение состояния заболеваний на хронические и острые с использованием модели RuRoberta

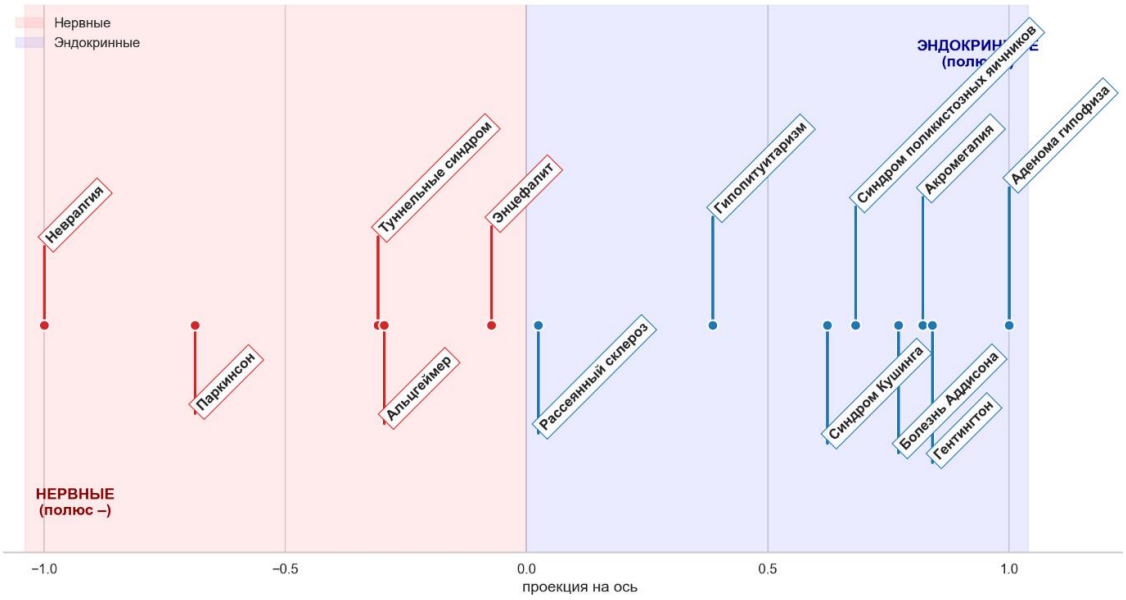

Рисунок 3 – Разделение состояния заболеваний на неврологические и эндокринные с использованием модели RuBioRoBERTa

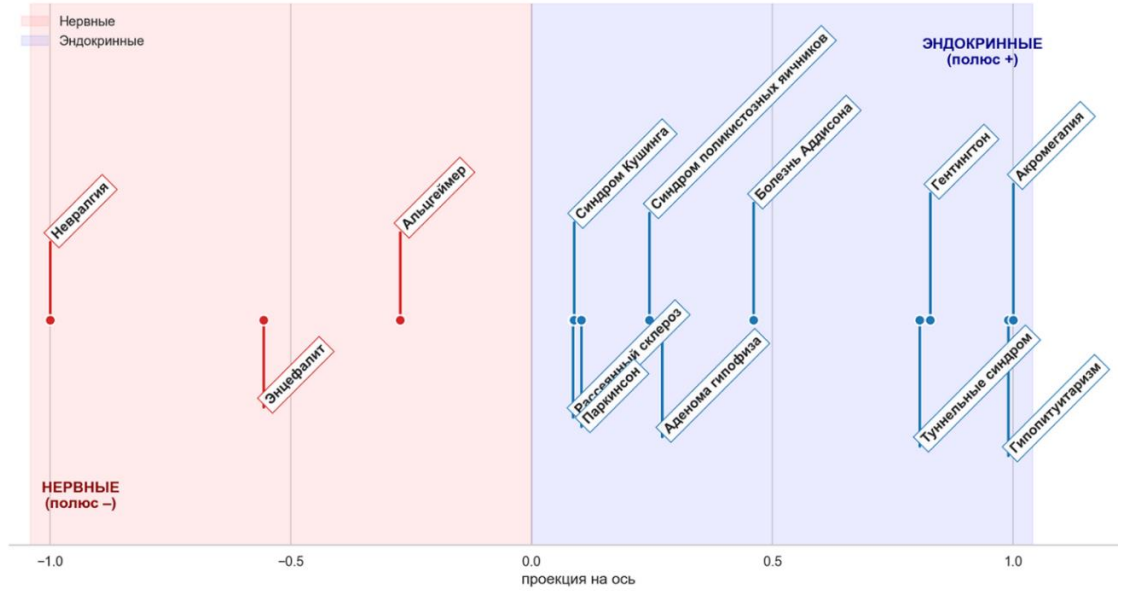

Рисунок 4 – Разделение состояния заболеваний на неврологические и эндокринные с использованием модели RuRoberta

model	mean_abs	std
Ось "Острое - Хроническое"		
RuBioRoBERTa	0.46	0.57
ruRoberta-large	0.47	0.53
Ось "Эндокринные - Неврологические"		
RuBioRoBERTa	0.58	0.62
ruRoberta-large	0.52	0.58

Рисунок 5 – Таблица сравнения обеих моделей

Результаты, представленные в соответствии с рисунками 1-4, подтверждают, что дообучение на узкоспециализированных корпусах не просто улучшает метрики на конкретных задачах, но и качественно меняет геометрию семантического пространства модели. RuBioRoBERTa эффективно понимает профессиональные медицинские термины.

При сравнительном анализе семантических осей “Острое–Хроническое” (в соответствии с рисунками 1, 2) выявлено существенное преимущество специализированной медицинской модели RuBioRoBERTa в точности воспроизведения клинической классификации. Модель формирует семантический континуум, в целом соответствующий медицинским представлениям. Однако наблюдаются отдельные случаи, требующие интерпретации.

Позиционирование менингита в области хронических заболеваний, хотя и противоречит его типичному острому течению, может семантически отражать существование хронических форм патологии (код G03.1 по МКБ-10). Аналогично, классификация аппендицита как хронического состояния потенциально учитывает возможность вялотекущего воспаления червеобразного отростка. В то же время отнесение холеры к хроническим заболеваниям, несмотря на её близость к нулевой точке проекции, следует признать семантической ошибкой, вероятно, обусловленной недостаточной репрезентативностью данного понятия в обучающем корпусе.

Критически важным является некорректное позиционирование острых неотложных состояний – субарахноидального кровоизлияния и транзиторной ишемической атаки – в зоне хронических заболеваний общей моделью ruRoberta. В сценарии использования модели для определения приоритетности медицинской помощи такая ошибка могла бы иметь серьезные практические последствия.

Особый интерес представляет согласованное отнесение рака молочной железы к острым состояниям обеими моделями, что противоречит его объективно хроническому характеру течения. Данный семантический сдвиг, вероятно, обусловлен доминированием в текстовых корпусах контекстов, связанных с активной фазой диагностики и неотложного начала лечения, что акцентирует «остроту» реакции на заболевание, а не его естественную историю.

Таким образом, специализированная модель не только демонстрирует более высокую общую объяснимость расположения диагнозов, но и критически важное — с прикладной точки зрения — преимущество в корректной семантической классификации острых, угрожающих жизни состояний.

Высокое стандартное отклонение (0.57) говорит о том, что модель не всегда уверенно относит конкретное заболевание к одному из полюсов, что может отражать реальную клиническую неоднозначность. Таким образом, "неуверенность" модели может быть не недостатком, а отражением амбивалентности самих понятий.

Анализ оси «Эндокринные–Неврологические» (в соответствии с рисунками 3, 4) подтверждает преимущество доменно-адаптированной модели. RuBioRoBERTa корректно классифицирует 5 из 7 неврологических заболеваний против 3 у ruRoberta-large, при равной эффективности в определении эндокринных патологий.

Медицинская модель демонстрирует более тонкое понимание пограничных состояний: она точнее позиционирует рассеянный склероз ближе к неврологическому полюсу, тогда как общая модель верно определяет междисциплинарный характер аденомы гипофиза, относя её ближе к центру оси. Обе модели демонстрируют ошибку в классификации болезни Гентингтона. Это может быть связано с данными, на которых обучались трансформеры.

Таким образом, доменная адаптация позволяет модели не только точнее классифицировать профильные заболевания, но и формировать семантические представления, более адекватно отражающие клиническую реальность, включая нюансы междисциплинарных патологий.

Количественная оценка (в соответствии с рисунком 5) показывает, что RuBioRoBERTa достигает более высоких значений средней абсолютной проекции при меньшем стандартном отклонении, что свидетельствует о более уверенном и структурно организованном семантическом пространстве. В отличие от «размытой» картины общей модели, специализированная демонстрирует четкую семантическую дифференциацию, критически важную для медицинских приложений. Таким образом, доменная адаптация не только улучшает формальные метрики, но и повышает клиническую релевантность векторных представлений.

Выводы

Библиотека Semgeom для языка Python, разработанная авторами, результаты обработки моделей которой, представлены в этой статье, размещена онлайн и общедоступна [9].

Она продемонстрировала высокую эффективность как инструмент для анализа семантических пространств языковых моделей, позволяя:

- сравнивать различные модели по степени отражения доменных знаний и структурировать их семантическое пространство;
- валидировать предметную адекватность эмбеддингов, выявляя насколько проекции терминов соответствуют экспертным представлениям о понятиях;
- определять слабые места моделей, которые требуют дополнительного дообучения на специализированных корпусах.

Эксперимент с русскоязычными моделями RuBioRoBERTa и ruRoberta-large показал, что доменно-адаптированная медицинская модель формирует семантическое пространство, более точно отражающее профессиональные противопоставления. А высокие значения среднего абсолютного значения проекций и умеренное стандартное отклонение проекций у RuBioRoBERTa указывают на более чёткое разделение понятий и согласованность ассоциаций терминов.

Результаты могут быть использованы для выбора модели при разработке систем поддержки принятия медицинских решений. Сама методика позволяет выявлять слабые места модели и целенаправленно корректировать обучение на специфических доменных данных. Визуализация и количественные метрики обеспечивают инструментарий для интерпретируемой оценки LLM в профессиональных приложениях, снижая риск ошибок при работе с критически важной информацией.

Список литературы

1. Attention is all you need / Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. //Advances in neural information processing systems. 2017. Vol. 30. P.1-11.
2. Imparting interpretability to word embeddings while preserving semantic structure / Senel L.K., Utlu ¸ I., ¸Sahinuç F., Ozaktas H.M., Koç A. // Natural Language Engineering. 2021. Vol. 27. No. 6. P. 721–746.
3. A survey of the state of explainable AI for natural language processing / Marina Danilevsky M., Qian K., Aharonov R., Katsis Y., Kawas B., Sen P.// Proceedings of AACL-IJCNLP 2020. 2020. P.1-13
4. Concept-based explainable artificial intelligence: A survey/ Poeta E., Ciravegna G., Pastor E., Cerquitelli T., Baralis E. // ACM Computing Surveys. 2025. P.1-42
5. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? / Huang G., Li Y., Jameel S., Long Y., Papanastasiou G. //Computational and structural biotechnology journal. 2024. Vol. 24. P. 362-373.
6. Semantic structure in large language model embeddings / Kozlowski A. C., Dai C., Boutyline A. // URL: arXiv preprint arXiv:2508.10003. 2025. (Дата обращения: 31.08.2025)
7. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment / An J., Kwak H., Ahn Y. Y. // Proceedings of ACL 2018. 2018. P. 1-12.
8. Mathematical features of semantic projections and word embeddings for automatic linguistic analysis / de Córdoba P. F., Pérez C. A. R., Pérez E. A. S. //AIMS MATHEMATICS. 2025. Vol. 10. No. 2. P. 3961-3982.
9. Библиотека Semgeom 1.1 / Топпер А. // [Электронный ресурс] URL: https://pypi.org/project/semgeom/ (Дата обращения: 31.08.2025)

10. On convex decision regions in deep network representations/ Tětková L., Brüsch T., Dorszewski T. *et al.* //Nature Communications. 2025.Vol. 16. No. 1. P. 5419.
11. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings / Grand G., Blank IA., Pereira F., Fedorenko E. // URL: arXiv preprint arXiv:1802.01241. 2018. (Дата обращения: 31.01.2025)
12. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings / Mathew B.,Sikdar S., Lemmerich F., Strohmaier M. // Proceedings of The Web Conference 2020. 2020. P. 1548–1558.
13. SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings / Engler J., Sikdar S., Lutz M., Strohmaier M. // Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. P. 4607–4619.
14. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining / Yalunin A., Nesterov A., Umerenkov D. // URL: arXiv preprint arXiv:2204.03951. 2022. (Дата обращения: 31.01.2025)
15. A family of pretrained transformer language models for Russian/ Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A. //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024. P. 507-524.
16. Towards Domain Specification of Embedding Models in Medicine / Khodadad M., Kasmaee AS., Astaraki M., Mahyar H. // URL: arXiv preprint arXiv:2507.19407. 2025. (Дата обращения: 31.01.2025)
17. Discovering universal geometry in embeddings with ICA / Yamagiwa H., Oyama M., Shimodaira H. // URL: arXiv preprint arXiv:2305.13175. 2023. (Дата обращения: 31.01.2025)
18. Axis tour: Word tour determines the order of axes in ica-transformed embeddings / Yamagiwa H., Takase Y., Shimodaira H. // URL: arXiv preprint arXiv:2401.06112. 2024. (Дата обращения: 31.01.2025)
19. A systematic review of bias detection methods for non-English word embeddings and language models/ Puttick A., Ikae1 C., Rigotti C., Fosch-Villaronga E., Kharas M.W., Søraa R.A., Kurpicz-Briki M. // Artificial Intelligence Review. 2025. Vol. 58. No. 12. P. 1–56.
20. Visualization of medical concepts represented using word embeddings: a scoping review / Oubenali N., Messaoud S., Filiot A., Lamer A., Andrey P. // BMC Medical Informatics and Decision Making. 2022. Vol. 22. No. 1. P. 83.
21. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases / Chen Z., Liu, X., Bian J. // BMC Medical Informatics and Decision Making. 2018. Vol. 18. No. S2. P. 65.
22. Визуальное мышление в виртуальном мире управления и принятия решений / Колесников А. В. Листопад С.В., Бенько А.И., Майтаков Ф.Г. // Проблемы искусственного интеллекта. 2017. № 4 (7). С. 49–59.
23. Система автоматической адаптации русскоязычных текстов и ее практическая значимость / Большакова С. А. // Проблемы искусственного интеллекта. 2024. Т. 34. № 3. С. 45–54.
24. CSpace: A concept embedding space for bio-medical applications / Tomasoni D., Marchetti L. C. // Bioinformatics. 2025. P. btaf376.

References

1. Attention is all you need / Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. //Advances in neural information processing systems. 2017. Vol. 30. P.1-11.
2. Imparting interpretability to word embeddings while preserving semantic structure / Senel L.K., Utlu ¸ I., ¸Sahinuç F., Ozaktas H.M., Koç A. // Natural Language Engineering. 2021. Vol. 27. No. 6. P. 721–746.
3. A survey of the state of explainable AI for natural language processing / Marina Danilevsky M., Qian K., Aharonov R., Katsis Y., Kawas B., Sen P.// Proceedings of AACL-IJCNLP 2020. 2020. P.1-13
4. Concept-based explainable artificial intelligence: A survey/ Poeta E., Ciravegna G., Pastor E., Cerquitelli T., Baralis E. // ACM Computing Surveys. 2025. P.1-42
5. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? / Huang G., Li Y., Jameel S., Long Y., Papanastasiou G. //Computational and structural biotechnology journal. 2024. Vol. 24. P. 362-373.
6. Semantic structure in large language model embeddings / Kozlowski A. C., Dai C., Boutyline A. // URL: arXiv preprint arXiv:2508.10003. 2025. (accessed: 31.08.2025)
7. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment / An J., Kwak H., Ahn Y. Y. // Proceedings of ACL 2018. 2018. P. 1-12.
8. Mathematical features of semantic projections and word embeddings for automatic linguistic analysis / de Córdoba P. F., Pérez C. A. R., Pérez E. A. S. //AIMS MATHEMATICS. 2025. Vol. 10. No. 2. P. 3961-3982.
9. Semgeom Library 1.1 / Topper A. // [Electronic resource]. Available at: URL: https://pypi.org/project/semgeom/ (accessed 31.08.2025).
10. On convex decision regions in deep network representations/ Tětková L., Brüsch T., Dorszewski T. *et al.* //Nature Communications. 2025.Vol. 16. No. 1. P. 5419.
11. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings / Grand G., Blank IA., Pereira F., Fedorenko E. // URL: arXiv preprint arXiv:1802.01241. 2018. (accessed: 31.01.2025)
12. The polar framework: Polar opposites enable interpretability of pre-trained word embeddings / Mathew B.,Sikdar S., Lemmerich F., Strohmaier M. // Proceedings of The Web Conference 2020. 2020. P. 1548–1558.

13. SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings / Engler J., Sikdar S., Lutz M., Strohmaier M. // Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. P. 4607–4619.
14. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining / Yalunin A., Nesterov A., Umerenkov D. // URL: arXiv preprint arXiv:2204.03951. 2022. (accessed: 31.01.2025)
15. A family of pretrained transformer language models for Russian/ Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A. //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024. P. 507-524.
16. Towards Domain Specification of Embedding Models in Medicine / Khodadad M., Kasmaee AS., Astaraki M., Mahyar H. // URL: arXiv preprint arXiv:2507.19407. 2025. (accessed: 31.01.2025)
17. Discovering universal geometry in embeddings with ICA / Yamagiwa H., Oyama M., Shimodaira H. // URL: arXiv preprint arXiv:2305.13175. 2023. (accessed: 31.01.2025)
18. Axis tour: Word tour determines the order of axes in ica-transformed embeddings / Yamagiwa H., Takase Y., Shimodaira H. // URL: arXiv preprint arXiv:2401.06112. 2024. (accessed: 31.01.2025)
19. A systematic review of bias detection methods for non-English word embeddings and language models/ Puttick A., Ikae1 C., Rigotti C., Fosch-Villaronga E., Kharas M.W., Søraa R.A., Kurpicz-Briki M. // Artificial Intelligence Review. 2025. Vol. 58. No. 12. P. 1–56.
20. Visualization of medical concepts represented using word embeddings: a scoping review / Oubenali N., Messaoud S., Filiot A., Lamer A., Andrey P. // BMC Medical Informatics and Decision Making. 2022. Vol. 22. No. 1. P. 83.
21. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases / Chen Z., Liu, X., Bian J. // BMC Medical Informatics and Decision Making. 2018. Vol. 18. No. S2. P. 65.
22. Visual thinking in the virtual world of control and decision-making / Kolesnikov A. V., Listopad S.V., Benko A.I., Maitakov F.G. // Problems of Artificial Intelligence. 2017. No. 4 (7). P. 49–59.
23. System for automatic adaptation of Russian-language texts and its practical significance / Bolshakova S. A. // Problems of Artificial Intelligence. 2024. Vol. 34. No. 3. P. 45–54.
24. CSpace: A concept embedding space for bio-medical applications / Tomasoni D., Marchetti L. C. // Bioinformatics. 2025. P. btaf376.

RESUME

Topper A. M., Goncharova A. B.

Improving The Structuring Of Concepts In Language Models Through Domain Adaptation

The rapid development of transformer-based language models has created an urgent need for methods to interpret their internal semantic representations, particularly in specialized domains such as medicine. Standard quality metrics do not reveal how well models capture domain-specific knowledge structures, creating a gap between machine learning performance and practical applicability in critical fields.

We developed the Semgeom library for analyzing the geometry of embedding spaces using semantic axes methodology. The study compared two Russian-language models: domain-adapted medical model and general domain model. We defined semantic axes relevant to medicine with corresponding pole words and calculated vector directions using pairwise difference averaging. Projections of medical terms were analyzed using mean absolute value and standard deviation metrics.

The domain-adapted model demonstrated superior performance in semantic structuring of medical concepts. It achieved higher mean absolute projection values and showed more clinically accurate classification of medical conditions. Critical errors were identified in the general model, including misclassification of emergency conditions as chronic diseases.

Domain adaptation qualitatively transforms the geometry of semantic spaces beyond simply improving task-specific metrics. The Semgeom library provides an effective tool for validating the domain adequacy of language models and identifying areas requiring additional training. The methodology has practical significance for developing reliable AI systems in medicine and other specialized domains where semantic accuracy is critical.

РЕЗЮМЕ

Топпер А.М., Гончарова А.Б.
Улучшение структуризации понятий в языковых моделях посредством доменной адаптации

Быстрое развитие языковых моделей на архитектуре трансформеров создало острую потребность в методах интерпретации их внутренних семантических представлений, особенно в специализированных областях, таких как медицина. Стандартные метрики качества не раскрывают, насколько хорошо модели отражают предметные знания, создавая разрыв между производительностью машинного обучения и практической применимостью в критически важных областях.

Авторами разработана библиотека Semgeom для анализа геометрии пространств эмбеддингов с использованием метода семантических осей. Исследование сравнивало две русскоязычные модели: доменно-адаптированная медицинская модель и общедоменная модель. Определены семантические оси, релевантные для медицины с соответствующими словами-полюсами и вычислили векторы направлений с использованием усреднения попарных разностей. Проекции медицинских терминов анализировались с помощью метрик среднего абсолютного значения и стандартного отклонения.

Доменно-адаптированная модель продемонстрировала превосходство в семантической структуризации медицинских понятий. Она достигла более высоких значений средних абсолютных проекций и показала более клинически точную классификацию медицинских состояний. В общей модели выявлены критические ошибки, включая классификацию неотложных состояний как хронических заболеваний.

Доменная адаптация качественно преобразует геометрию семантических пространств, выходя за рамки простого улучшения метрик на конкретных задачах. Библиотека Semgeom предоставляет эффективный инструмент для валидации предметной адекватности языковых моделей и выявления областей, требующих дополнительного обучения. Методология имеет практическое значение для разработки надежных систем ИИ в медицине и других специализированных областях, где семантическая точность имеет критическое значение.

Топпер Алина Михайловна – магистр 2 курса, Санкт-Петербургский государственный университет, программа магистратуры: прикладная математика и информатики в задачах медицинской диагностики, e-mail: st089228@student.spbu.ru. ORCID: 0009-0000-3313-5399. Область научных интересов: языковые модели, семантический анализ, искусственный интеллект в медицине.

Гончарова Анастасия Борисовна – к.ф.-м.н., доцент, Санкт-Петербургский государственный университет, кафедра Теории систем управления электрофизической аппаратурой, e-mail: a.goncharova@spbu.ru. ORCID: 0000-0002-7980-1657. Область научных интересов: системы поддержки принятия решений для медицины, обработка медицинских данных, математическое моделирование в медицине, машинное обучение и ИИ.

Topper Alina Mikhailovna - 2nd-year Master's student at St. Petersburg State University, majoring in Applied Mathematics and Computer Science in Medical Diagnostics, e-mail: st089228@student.spbu.ru. ORCID: 0009-0000-3313-5399. Research interests: include language models, semantic analysis, and artificial intelligence in medicine.

Goncharova Anastasia Borisovna – PhD, Associate Professor, St. Petersburg State University, Department of Theory of Control Systems for Electrophysical Equipment, e-mail: a.goncharova@spbu.ru. ORCID: 0000-0002-7980-1657. Research interests: decision support systems for medicine, medical data processing, mathematical modeling in medicine, machine learning and AI.

Статья поступила в редакцию 19.12.2025