Н

УДК 004.8+004.912 **DOI** 10.24412/2413-7383-2025-4-39-229-240

К. А. Никитенко, А. В. Звягинцева
ФГБОУ ВО «Донецкий государственный университет»
283001, г. Донецк, ул. Университетская, 24

ОЦЕНКА СЕМАНТИЧЕСКОЙ СХОЖЕСТИ ПРЕДЛОЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ ВЕРОЯТНОСТНЫХ МЕР ЭМБЕДДИНГОВ

K. A. Nikitenko, A. V. Zviagintseva
Federal State Educational Institution of Higher Education «Donetsk State University»
283001, Donetsk, University str, 24

ASSESSMENT OF SEMANTIC SIMILARITY OF SENTENCES USING PROBABILISTIC MEAUSRES BASED ON EMBEDDINGS

К. А. Нікітенко, Г. В. Звягінцева
ФДБОУ ВО «Донецький державний університет»
283001, м. Донецьк, вул. Університетська, 24

ОЦІНКА СЕМАНТИЧНОЇ СХОЖОСТІ РЕЧЕНЬ З ВИКОРИСТАННЯМ ІМОВІРНІСНИХ МІР ЕМБЕДДИНГІВ

В статье рассматривается задача оценки семантической схожести предложений с использованием вероятностных мер эмбеддингов слов. Предложен метод, основанный на сопоставлении вероятностей независимых и зависимых событий, соответствующих словам в предложении. Проведен анализ методов построения матрицы внимания и оцифровки слов, обоснован выбор вероятностных эмбеддингов как основы для количественного описания смысловых зависимостей. Предложена процедура сопоставления вероятностей, вычисленных на основе эмбеддингов и весов внимания, что позволяет формализовать смысловые связи между словами и предложениями.
Ключевые слова: семантическая схожесть, вероятностная модель, матрица внимания, эмбеддинг, вероятностный анализ текста.

The article addresses the problem of evaluating the semantic similarity of sentences using probabilistic measures of word embeddings. An method is proposed based on comparing the probabilities of independent and dependent events corresponding to the words within a sentence. The methods for constructing attention matrices and digitizing words are analyzed, and the choice of probabilistic embeddings as the core of the semantic representation is justified. A procedure is proposed for comparing probabilities derived from embeddings and attention weights, which enables formalization of semantic relations between words and sentences.
Key words: semantic similarity, probabilistic model, attention matrix, embedding, probabilistic text analysis.

У статті розглядається задача оцінки семантичної схожості речень з використанням імовірнісних мір ембеддингів слів. Запропоновано метод, який засновано на зіставленні ймовірностей незалежних і залежних подій, що відповідають словам у реченні. Проведено аналіз методів побудови матриці уваги та оцифровки слів, обґрунтовано вибір імовірнісних ембеддингів як основи для кількісного опису смислових залежностей. Запропоновано процедуру зіставлення ймовірностей, які обчислено на основі ембеддингів та ваг уваги, що дозволяє формалізувати смислові зв'язки між словами та реченнями.
Ключові слова: семантична схожість, імовірнісна модель, матриця уваги, ембеддинг, імовірнісний аналіз тексту.

Введение

Современные методы анализа естественного языка всё активнее используют вероятностные и геометрические модели представления текстовой информации. Наиболее распространенные инструменты семантического анализа – распределенные представления слов, а также механизмы внимания, лежащие в основе архитектуры трансформеров [1]. Эти подходы обеспечивают высокие результаты в задачах машинного перевода, текстовой классификации, поиска и генерации ответов, однако сохраняют ряд теоретических и интерпретационных ограничений [2].

Одной из актуальных проблем современной лингвистической информатики является **оценка семантической схожести предложений,** то есть количественное измерение степени смысловой близости между высказываниями [3], [4]. В классических моделях распределенного представления слов, таких как Word2Vec, GloVe или FastText, семантические связи выражаются через геометрическую близость векторных представлений. Однако подобная близость не всегда имеет прямое вероятностное толкование. В то же время механизм внимания в трансформерах формирует матрицу весов, нормализованных по функции *softmax,* которые могут быть интерпретированы как вероятности условной зависимости между словами в предложении. Таким образом, возникает возможность сопоставления двух вероятностных пространств: одного, заданного эмбеддингами слов, и другого, определяемого распределением внимания.

В данной работе предлагается метод, основанный на сравнении **независимых** и **зависимых** вероятностных событий, соответствующих словам и их совместным появлениям в предложении. Если независимые события можно рассматривать как статистические соотнесения слов без смысловой связи, то зависимые отражают их реальную связанность в контексте. Такое сопоставление позволяет количественно оценить, насколько семантическая структура текста отклоняется от случайного распределения, и выявить закономерности смысловой организации предложений.

Цель исследования – разработка и тестирование вероятностного метода оценки семантической схожести предложений, использующего данные матрицы внимания и плотности распределений эмбеддингов слов.

Предлагаемый метод позволяет перейти от качественной интерпретации семантических связей к их **количественной вероятностной оценке,** что соответствует вероятностной природе смысловых отношений между языковыми единицами [5].

Обзор методов построения матрицы внимания и оцифровки слов

Механизм внимания является ключевым компонентом современных нейросетевых архитектур, предназначенных для обработки текстов. Его основная идея заключается в том, что модель при анализе каждого слова учитывает влияние всех остальных слов в предложении, взвешивая их вклад в итоговое представление.

Современные подходы к построению матриц внимания включают несколько классов методов. Классический additive attention основан на обучении весовой функции, вычисляющей степень соответствия между состоянием декодера и входным вектором. Впоследствии Luong предложил multiplicative attention, обеспечивающий более эффективные вычисления за счёт скалярного произведения векторов. Метод scaled attention, лежащий в основе архитектуры Transformer, нормализует произведение на размерность скрытого пространства, что улучшает устойчивость при обучении [6].

Позднее были предложены модификации, направленные на сокращение вычислительной сложности и улучшение интерпретируемости – multi-head attention, sparse attention, linear attention, efficient attention, где матрица внимания оптимизируется для ускорения обработки длинных последовательностей.

Методы оцифровки слов также претерпели значительное развитие. К числу базовых относятся Word2Vec, GloVe и FastText, формирующие статические векторные представления слов. Более современные модели, такие как ELMo, BERT, RoBERTa, GPT, обеспечивают контекстно-зависимые представления, что делает возможным учёт смысла слова в конкретном контексте [7]. Для оценки семантической схожести предложений эффективными оказались модели Sentence-BERT [8] и SimCSE [9], использующие эмбеддинги на уровне предложений. Однако большинство из этих подходов не дают прямой вероятностной интерпретации, что делает актуальным развитие методов, основанных на плотностях распределений эмбеддингов [10].

Классическая формулировка механизма внимания впервые была предложена в работах Bahdanau и Luong для машинного перевода, а затем обобщена в архитектуре трансформеров, предложенных Vaswani. В трансформерах внимание реализуется через вычисление матрицы весов [11]:

$$a_{ij} = soft\max\left(\frac{\left(\left(x_i W_q\right)\cdot\left(x_j W_k\right)\right)^T}{\sqrt{d_k}}\right), \quad (1)$$

где a_{ij} – элемент матрицы внимания, показывающий насколько элемент x_i «внимателен» к элементу x_j; x_i, x_j – векторы эмбеддингов слов; W_q, W_k – обучаемые матрицы весов для формирования запроса и ключа соответственно; d_k – размерность вектора ключа.

Нормализация функции *softmax* обеспечивает выполнение условия $\sum_{j=1}^{n} a_{i,j} = 1$, что означает: каждая строка матрицы представляет вероятностное распределение, где сумма весов связей слова i со всеми словами предложения равна единице, и позволяет интерпретировать значения a_{ij} как вероятности зависимости между словами w_i и w_j.

Дальнейшее развитие механизма внимания привело к появлению многоголового внимания (multi-head attention), где информация о зависимостях извлекается в нескольких подпространствах признаков, что повышает точность, но затрудняет интерпретацию вероятностных весов.

Для задач вероятностного анализа текстов важно не столько высокое качество генерации, сколько интерпретируемость матрицы внимания. Поэтому в данной работе используется упрощённый вариант механизма самовнимания, основанный на нормализованном косинусном сходстве между векторами слов:

$$P\left(w_i \middle| w_j\right) = \frac{\exp\left(\frac{\cos\left(x_i x_j\right)}{T}\right)}{\sum_{k=1}^{n}\exp\left(\frac{\cos\left(x_i x_k\right)}{T}\right)}, \quad (2)$$

где $P\left(w_i \middle| w_j\right)$ – условная вероятность того, что слово w_i будет «связано» относительно слова w_j; x_i, x_j – векторные представления слов w_i и w_j соответственно в пространстве признаков; T – параметр «температуры», регулирующий «резкость» распределения; k – порядковый номер слова в анализируемой выборке текста.

Такой способ позволяет получить матрицу вероятностей без обучения модели и сохранить её статистическую интерпретацию. Каждая строка матрицы описывает распределение вероятностей связи данного слова с остальными словами предложения. Таким образом, матрица внимания рассматривается как вероятностная матрица условных зависимостей между словами, что делает её подходящим инструментом для последующего сравнения с вероятностями, вычисленными на плотностях эмбеддингов.

Методы оцифровки слов

Основой большинства моделей является гипотеза распределенного значения: слова, встречающиеся в похожих контекстах, имеют сходное значение. Наиболее известными методами векторного представления слов являются Word2Vec, GloVe и FastText. Эти методы создают точечные представления в пространстве высокой размерности, где расстояния между векторами отражают степень семантической близости слов. Современные модели, такие как ELMo, BERT, RoBERTa, формируют контекстно-зависимые представления слов: один и тот же лексемный элемент получает различные векторы в зависимости от окружения. Это повышает точность моделирования смысла, но усложняет вероятностную интерпретацию – эмбеддинги становятся функцией контекста, а не фиксированным распределением.

Для решения поставленной задачи целесообразно использовать вероятностные эмбеддинги, где каждое слово описывается гауссовым распределением [12]. Такой подход обеспечивает возможность оценивать вероятности появления отдельных слов, вычислять совместные вероятности групп слов (например, в пределах предложения) и сравнивать их с распределениями, полученными из матриц внимания. Это позволит рассматривать структуру текста как систему вероятностных зависимостей, где геометрические расстояния эмбеддингов связаны с вероятностями смысловых связей.

Описание исходных данных

Для экспериментальной проверки предложенного метода использовался корпус более чем из 20 произведений А.С. Пушкина общим объёмом 3105 строк разной длины, содержащий тексты в прозе и стихах, верифицированные по морфологической базе OpenCorpora [13]. Выбор корпуса обусловлен рядом факторов:

1) тексты отличаются лексическим и синтаксическим разнообразием, позволяющим анализировать зависимости между словами в разных типах предложений [14];

2) корпус удобно структурировать на отдельные предложения и токены, что облегчает формирование выборок для анализа;

3) литературный язык А.С. Пушкина хорошо представлен в существующих морфологических и лексических базах данных (таких как OpenCorpora и rusvectores), что обеспечивает корректную токенизацию и морфологическую разметку.

Метод анализа

Предложенный метод основан на сопоставлении вероятностей зависимых и независимых событий, вычисляемых на основе матрицы внимания [15]. Каждое слово рассматривается как элементарное вероятностное событие, а предложение – как их совокупность. На этой основе проводится сравнение распределений, полученных для независимого (частотного) и зависимого (контекстного) случаев, что позволяет оценить степень смысловой когерентности текста [16].

Особенность метода – использование упрощённого механизма самовнимания, где вероятности зависимостей между словами определяются через нормализованные косинусные меры сходства, а эмбеддинги трактуются как вероятностные распределения

в многомерном пространстве. Такой метод обеспечивает интерпретируемость полученных параметров и их статистическую сопоставимость [17]. Ключевая идея состоит в том, что каждое слово рассматривается как элементарное вероятностное событие, а предложение – как совокупность этих событий. Сравнение независимых и зависимых вероятностных событий проводится по одному и тому же набору слов.

Реализация и расчёты

Пусть корпус содержит m предложений, каждое из которых состоит из n_i слов. Для предложения $S_i = \{w_1, w_2, \dots, w_{n_i}\}$ формируется матрица вероятностей внимания размерности $n_i \times n_i$:

$$A_i = \left[P\left(w_i | w_j\right)\right], \tag{3}$$

где $P\left(w_i | w_j\right)$ – вероятность связи слова w_i с w_j, определяемая по нормализованным весам внимания.

Параллельно для каждого слова w_j строится вероятностное распределение плотности $p_j(x)$, полученное из выбранной гауссовой модели эмбеддингов. На основе распределений вычисляются одномерные и совместные вероятностные события.

Для иллюстрации результатов вычислений на корпусе из произведений А.С. Пушкина построены распределения вероятностных параметров и матрицы внимания.

Для анализа распределения вероятностных соотношений между зависимыми и независимыми событиями вычислялось значение натурального логарифма:

$$\ln \Delta = \ln \frac{P_{dep}}{P_{ind}}, \tag{4}$$

где P_{dep} – вероятность зависимого события, отражающая связь слова с контекстом, P_{ind} – вероятность независимого события (без учёта контекстных зависимостей). Значение lnΔ показывает степень смысловой связанности предложений: чем больше величина, тем сильнее выражены семантические зависимости между словами (рис. 1).

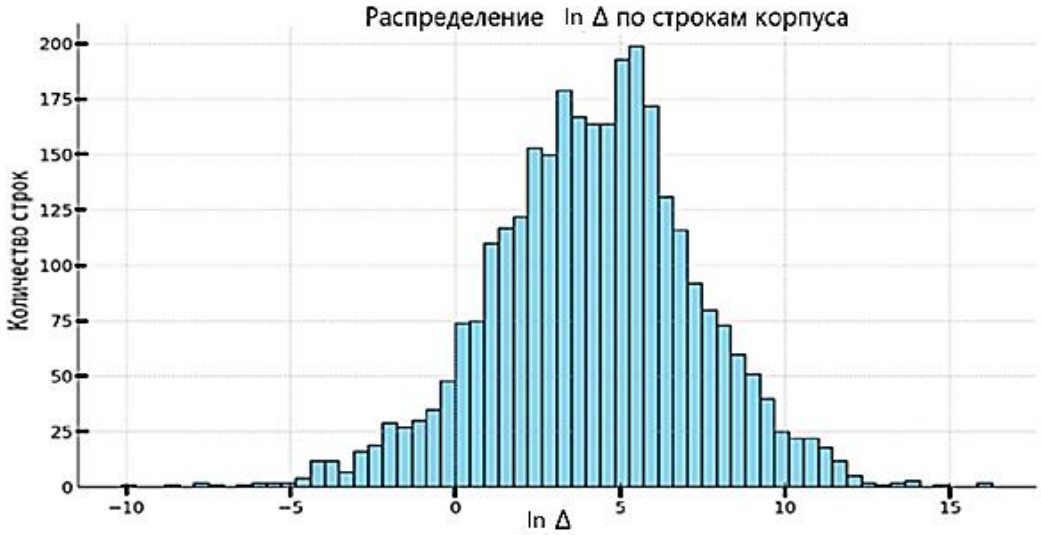

Рисунок 1 – Распределение lnΔ по строкам корпуса произведений А.С. Пушкина

Для более детального анализа рассмотрим строку корпуса «По нивам повлечет плуг, миром изощренный», извлечённую из поэмы А.С. Пушкина «Евгений Онегин». Данный фрагмент выбран, поскольку содержит типичную для литературного текста синтаксическую структуру с явно выраженными смысловыми зависимостями между словами, что позволяет наглядно проиллюстрировать работу предлагаемого метода.

Для каждого слова *w* в корпусе вычислим эмпирическую вероятность *P(w)*:

$$P(w) = \frac{f(w)}{N}, \tag{5}$$

где $f(w)$ – количество вхождений слова *w* в корпусе, *N* – общее число токенов (словоформ) в корпусе.

Величина $P(w)$ отражает относительную частоту появления конкретного слова в тексте и используется при оценке независимых вероятностей. Полученные значения для слов из рассматриваемого предложения приведены в таблице 1.

Таблица 1 – Вероятности появления слов в предложении в зависимости от других слов

Позиция	Слово	*P(w)*
1	По	0,002415
2	Нивам	0,000073
3	Повлечет	0,000073
4	Плуг	0,000146
5	Миром	0,000073
6	Изощренный	0,000073

Как упоминалось ранее, матрица внимания (3) построена для выбранной строки корпуса по выражению (2). В данном случае интерес представляют условные вероятности появления слов, вычисляемые в рамках цепочного приближения.

Для каждого слова w_j определяется вероятность его появления при условии предыдущего слова w_{j-1}:

$$P(w_j | w_{j-1}) = \frac{P(w_{j-1}, w_j)}{P(w_{j-1})}, \tag{6}$$

где $P(w_{j-1}, w_j)$ – совместная вероятность появления пары слов; $P(w_{j-1})$ – эмпирическая вероятность предыдущего слова.

Такие условные вероятности позволяют учесть локальные контекстные зависимости между словами внутри предложения. Для рассматриваемой строки корпуса получены результаты, приведенные в таблице 2.

Таблица 2 – Условные вероятности для слов в заданном предложении

Позиция *j*	Слово w_j	$P(w_j \mid w_{j-1})$
2	Нивам	0,000612
3	Повлечет	0,003428
4	Плуг	0,001514
5	Миром	0,009611
6	Изощренный	0,001610

Для позиции 1 («По») значения предыдущего слова отсутствует, поэтому условная вероятность не определяется и в таблицу не включается.

При предположении статистической независимости слов предложение S можно описать как произведение индивидуальных вероятностей входящих в него слов:

$$P_{ind}(S) = \sum_{j=1}^{n} P(w_j). \quad (7)$$

Подставляя значения из таблицы 1, получаем:

$P_{ind}(S) = 0{,}002415 \cdot 0{,}000073 \cdot 0{,}000073 \cdot 0{,}000146 \cdot 0{,}000073 \cdot 0{,}000073 \approx 1{,}01 \cdot 10^{-23}$.

В логарифмическом масштабе: $\ln P_{ind} \approx -52{,}944586$.

Для зависимого (цепного) случая вероятность предложения вычисляется с учётом условных вероятностей из таблицы 2:

$$P_{dep}(S) = P(w_1) \cdot \prod_{j=2}^{n} P(w_j | w_{j-1}). \quad (8)$$

Подставляя значения, получаем:

$P_{dep}(S) = P(w_1) \cdot \times \sum_{j=2}^{6} P(w_j | w_{j-1}) = 0{,}002415 \cdot 0{,}000612 \cdot 0{,}003428 \cdot 0{,}001514 \cdot 0{,}009611 \cdot 0{,}001610 \approx 1{,}19 \cdot 10^{-16}$.

В логарифмическом виде: $\ln P_{dep} \approx -36{,}670714$.

Разница между зависимой и независимой вероятностями характеризует степень семантической связанности слов в предложении:

$\ln\Delta = \ln P_{dep} - \ln P_{ind} = -36{,}67 - (-52{,}94) = 16{,}27$.

Тогда отношение вероятностей составляет:

$\frac{P_{dep}}{P_{ind}} = \approx e^{16{,}273872} \approx 1{,}17 \cdot 10^{7}$.

Таким образом, зависимая вероятность рассматриваемой строки примерно в 11,7 миллионов раз выше, чем вероятность того же набора слов при условии их статистической независимости. Полученное значение $\ln\Delta \approx 16{,}27$ является высоким положительным показателем, что свидетельствует о сильной смысловой связанности слов в данном предложении.

Матрица размером 6 × 6 (рис. 2) визуализирует нормализованные вероятности условных событий (3). В рассматриваемом примере предложение «по нивам повлечет плуг миром изощренный» содержит $n = 6$ слов, для которых матрица внимания отражает степень вероятностной зависимости каждого слова от отдельных. Ярко выраженные значения на диагонали ($P(w_j) \approx 1$) соответствуют самовниманию – максимальной значимости слова относительно себя. Низкие значения вне диагонали ($P(w_j) \ll 1$) фиксируют слабые связи между различными словами. Таким образом, матрица внимания служит вероятностной аппроксимацией смысловых отношений внутри предложения и позволяет количественно оценить распределение внутренней связи текста на уровне слов.

Для расчёта вероятностей зависимых и независимых событий, отражающих связь слова с контекстом, а также определения степени смысловой связанности предложений разработан скрипт. Для примера в таблице 3 представлены строки корпуса, для которых отношения зависимых и независимых вероятностей принимают наибольшие значения. Эти предложения характеризуются наибольшей семантической когерентностью, что подтверждается высокими положительными значениями $\ln\Delta$.

В таблице 3 приведены длины предложений (n) и рассчитанные значения $\ln(P_{ind})$, $\ln(P_{dep})$ и $\ln\Delta$. Максимальное значение $\ln\Delta$=16,27 получено для строки «По нивам повлечет плуг миром изощренный», что соответствует примерно $1{,}17 \cdot 10^{7}$ – кратному превышению вероятности зависимой модели над независимой.

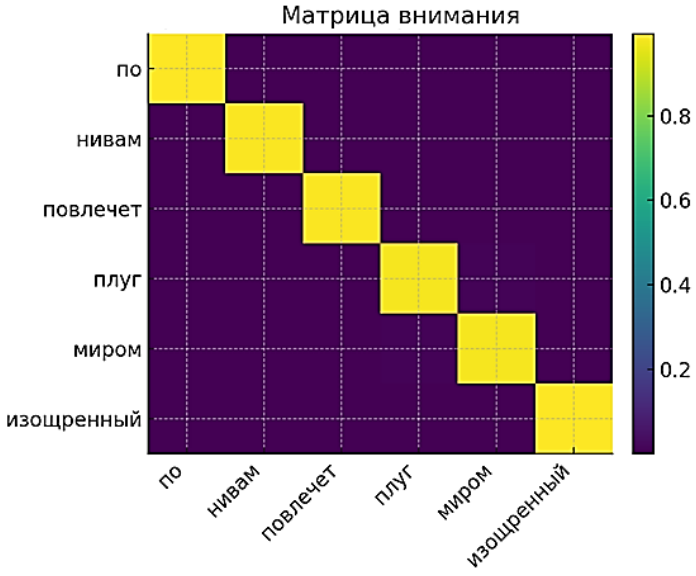

Рисунок 2 – Матрица внимания для выбранной строки

Таблица 3 – Предложения с наибольшей семантической связанностью по значению lnΔ

№	Предложение	Длина (n)	$\ln(P_{ind})$	$\ln(P_{dep})$	lnΔ
1	По нивам повлечет плуг, миром изощренный	6	-52,945	-36,671	16,274
2	И громко всем кричал: «Нашел! нашел!»	6	-44,301	-28,376	15,925
3	Держись, держись всегда прямой дороги	5	-43,230	-28,380	14,849
4	И много, много сильных пало	5	-38,468	-24,513	13,955
5	Одной, одной Осгар Мальвиною дышал	5	-42,760	-28,931	13,829
6	Из трубки пенковой дым гонит сероватый	6	-51,723	-38,060	13,663
7	В гордыне возмечтав мечом низвергнуть троны	6	-50,280	-36,902	13,378
8	На лире б возгремел гармонией небесной	6	-47,133	-33,838	13,295
9	Ни чистым золотом набиты сундуки	5	-44,973	-32,035	12,938
10	И месяц, дальних туч покинув темны сени	7	-56,596	-43,863	12,733

Представленные предложения имеют наибольшую семантическую связанность, поскольку их зависимые вероятности значительно превышают независимые, что подтверждается высокими значениями lnΔ. Во всех случаях наблюдается ярко выраженная структурная и смысловая целостность выражений, в том числе за счёт повторов и устойчивых словосочетаний. Это показывает, что предложенный метод эффективно выявляет фрагменты текста с плотными семантическими связями.

Обобщение результатов практических расчётов

Проведенные вычисления на корпусе произведений А.С. Пушкина подтвердили эффективность предложенного вероятностного метода оценки семантической связности предложений. Сравнение независимых и зависимых событий показало, что логарифмическая разность их вероятностей распределена близко к нормальному закону, что свидетельствует о статистической устойчивости метода. Наибольшие значения данного показателя фиксируются в предложениях с выраженными смысловыми и грамматическими зависимостями, где наблюдается высокая плотность контекстуальных связей между словами. Таким образом, практические расчёты демонстрируют способность метода адекватно отражать смысловую когерентность текста на уровне предложений.

Матрица внимания позволила формализовать взаимодействие слов в пределах предложения и визуализировать вероятностные зависимости между ними, что сделало возможным количественное описание структуры контекста. В то же время применение вероятностных эмбеддингов обеспечило переход от дискретных частотных характеристик к непрерывным распределениям, отражающим семантическую «плотность» слов в пространстве смыслов. Совместное использование этих инструментов дало возможность оценить как локальные связи внутри предложений, так и их глобальную смысловую организацию.

В целом результаты расчётов подтверждают, что интеграция вероятностных представлений слов и механизмов внимания обеспечивает достоверное количественное измерение семантической связанности предложений.

Выводы

В результате проведенного исследования разработан и реализован вероятностный метод оценки семантической связности предложений, основанный на сопоставлении независимых и зависимых событий в пространстве эмбеддингов. Применение матрицы внимания позволило перейти от частотного описания слов к вероятностному моделированию их взаимного влияния в контексте, что обеспечивает более глубокое понимание смысловой структуры текста.

Экспериментальная проверка метода на корпусе произведений А.С. Пушкина показала, что логарифмическая разность вероятностей зависимых и независимых событий подчиняется распределению, близкому к нормальному. Высокие значения этой величины соответствуют предложениям с выраженной смысловой когерентностью, что подтверждает корректность предложенного вероятностного метода к оценке внутренней связанности текста.

Научная новизна работы заключается в объединении вероятностного анализа эмбеддингов и механизмов внимания для совместного описания семантических зависимостей. В отличие от традиционных метрик, таких как косинусная близость или частотные показатели, предложенный метод вводит интерпретируемые вероятностные параметры [18], позволяющие количественно оценить переход от независимых слов к зависимым контекстам и представить смысловую структуру текста в вероятностно-геометрических терминах.

Практическая значимость результатов определяется возможностью их применения в задачах:

- автоматической оценки когерентности текста и качества генерации в нейросетевых моделях;
- семантического поиска и кластеризации предложений по смысловой близости;
- выявления аномальных или малоинформативных фраз в корпусах;
- предварительной фильтрации текстовых данных для обучения языковых моделей.

Перспективы дальнейших исследований связаны с расширением методики для анализа масштабных корпусов и интеграцией с современными трансформерными архитектурами (BERT, GPT), что позволит уточнить вероятностные зависимости на уровне контекстуальных эмбеддингов. Полученные результаты подтверждают, что вероятностные меры, основанные на матрицах внимания, способны выявлять скрытую структуру текста и формализовать понятие смысловой связности в терминах сопоставления зависимых и независимых событий, что делает предложенный метод перспективным инструментом семантического анализа и интерпретации моделей.

Список литературы

1. Minaee S., Kalchbrenner N., Cambria E., Nikzad N.,Chenaghlu M., Gao J. Deep Learning-Based Text Classification: A comprehensive review // ACM Computing Surveys. no 54(3), 2021: 1–40.
2. Аверин Г.В. О вероятностной природе смыслов в дискретных языковых единицах // Системный анализ и информационные технологии в науках о природе и обществе. №1(12)–2(13), 2017. С. 11–18.
3. Андриевская Н.К. Гибридная интеллектуальная мера оценки семантической близости // Проблемы искусственного интеллекта. №1(20), 2021. С. 4–17.
4. Меры семантической близости в онтологии / К.В. Крюкова, Л.А. Панкова, В.А. Пронина, В.С. Суховеров, Л.Б. Шипилина // Проблемы управления. Вып. 5, 2010. С. 2–14.
5. Erk K. The probabilistic turn in semantic and pragmatics // Annu. Rev. Linguist. 2022. 8:101–21.
6. Zheng Z., Wang Y., Huang Y., Song S., Yang M., Tang B., Xiong F., Li Z., Attention Heads of Large Language Models: A Survey // arXiv – 2024 – arXiv:2409.03752.
7. Собчишен А.С., Звягинцева А.В. Вероятностно-смысловые модели оцифрованных текстовых данных // Материалы конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками». №9, 2024. – С. 55–59.
8. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // arXiv 2019 arXiv:1908.10084.
9. Gao T., Yao X., Chen D. SimCSE: Simple Constrastive Learning of Sentence Embeddings // arXiv – 2021 arXiv:2104.08821.
10. Shen L., Jiang H., Liu L., Shi S. Sen2Pro: A Probabilistic Perspective to Sentence Embedding from Pre-trained Language Models // arXiv – 2023 – arXiv:2306.02247.
11. Vaswani A., Shazeer N. Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need // arXiv – 2017 – arXiv:1706.03762.
12. Yoda S., Tsukagoshi H., Sasano R., Takeda K. Sentence Representations via Gaussian Embedding // arXiv – 2023 – arXiv:2305.12990v2.
13. Сайт проекта «Открытый корпус» (OpenCorpora) русского языка. – Электр. рес. – URL: https://opencorpora.org/ (01.11.2025).
14. Chun S., Joon S., Sampaio R., Kalantidis Y., Larlus D. Probabilistic Embeddings for Cross-Modal Retrieval // arXiv – 2021 – arXiv:2101.05068.
15. Abdelwahab A., Landwehr N. Deep distributional sequence embeddings based on a Wasserstein loss // arXiv – 2019 – arXiv:1912.01933.
16. Botha J. Probabilistic modeling of morphologically rich languages // arXiv – 2015 – arXiv:1508.04271.
17. Бондаренко В.И., Елисеев В.О., Ермоленко Т.В. Анализ эффективности глубоких языковых моделей для задачи определения тональности русскоязычных текстов // Проблемы искусственного интеллекта. №1(32), 2024. – С. 51–62.
18. Никитенко К.А, Звягинцева А.В. Интерпретируемость нейросемантических моделей при их применении в прикладных областях // Проблемы искусственного интеллекта. №2(37), 2025. С. 79–90.

References

1. Minaee S., Kalchbrenner N., Cambria E., Nikzad N., Chenaghlu M., Gao J. Deep Learning-Based Text Classification: A comprehensive review // ACM Computing Surveys. no 54(3), 2021: 1–40.
2. Averin G.V. O veroyatnostnoi prirode smyslov v diskretnykh yazykovykh edinitsakh [On the Probabilistic Nature of Meanings in Discrete Linguistic Units]. Sistemnyi analiz i informatsionnye tekhnologii v naukakh o prirode i obshchestve. no 1(12)–2(13), 2017: 11–18.
3. Andrievskaya N.K. Gibridnaya intellektual'naya mera ocenki semanticheskoj blizosti // Problemy iskusstvennogo intellekta. №1(20), 2021: 4–17.
4. Kryukova K.V., L.A. Pankova, V.A. Pronina, V.S. Sukhoverov, L.B. Shipilina. Mery semanticheskoy blizosti v ontologii [Measures of semantic proximity in ontology] Probl. upravl., 2010. Issue 5: 2–14.
5. Erk K., The probabilistic turn in semantics and pragmatics // Annu. Rev. Linguist. 2022. 8:101–21.
6. Zheng Z., Wang Y., Huang Y., Song S., Yang M., Tang B., Xiong F., Li Z., Attention Heads of Large Language Models: A Survey // arXiv – 2024 – arXiv:2409.03752.
7. Sobchishen A.S., Zviagintseva A.V. Veroyatnostno-smyslovye modeli otsifrovannykh tekstovykh dannykh [Probabilistic-Semantic Models of Digitized Text Data]. Materialy konferentsii “Matematicheskoe i komp'yuternoe modelirovanie v ekonomike, strakhovanii i upravlenii riskami”. no 9, 2024: 55–59.

8. Reimers N., Gurevych I., Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // arXiv – 2019 – arXiv:1908.10084.
9. Gao T., Yao X., Chen D. SimCSE: Simple Constrastive Learning of Sentence Embeddings // arXiv – 2021 – arXiv:2104.08821.
10. Shen L., Jiang H., Liu L., Shi S. Sen2Pro: A Probabilistic Perspective to Sentence Embedding from Pre-trained Language Models // arXiv – 2023 – arXiv:2306.02247.
11. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I., Attention is all you need // arXiv – 2017 – arXiv:1706.03762.
12. Yoda S., Tsukagoshi H., Sasano R., Takeda K. Sentence Representations via Gaussian Embedding // arXiv – 2023 – arXiv:2305.12990v2.
13. Sait proekta “Otkrytyi korpus” (OpenCorpora) russkogo yazyka [Project website “OpenCorpora” of the Russian language]. – Elektronnyi resurs. URL: https://opencorpora.org/ (accessed November 1, 2025).
14. Chun S., Joon S., Sampaio R., Kalantidis Y., Larlus D. Probabilistic Embeddings for Cross-Modal Retrieval // arXiv – 2021 – arXiv:2101.05068.
15. Abdelwahab A., Landwehr N. Deep distributional sequence embeddings based on a Wasserstein loss // arXiv – 2019 – arXiv:1912.01933.
16. Botha J. Probabilistic modeling of morphologically rich languages // arXiv – 2015 – arXiv:1508.04271.
17. Bondarenko V.I., Eliseev V.O., Ermolenko T.V. Analiz effektivnosti glubokikh yazykovykh modelei dlya zadachi opredeleniya tonal'nosti russkoyazychnykh tekstov [Analysis of the Effectiveness of Deep Language Models for Sentiment Analysis of Russian Texts]. Problemy iskusstvennogo intellekta. no 1(32), 2025: 51–62.
18. Nikitenko K.A., Zviagintseva A.V. Interpretiruemost' neirosemanticheskikh modelei pri ikh primenenii v prikladnykh oblastyakh [Interpretability of Neurosemantic Models in Applied Domains]. Problemy iskusstvennogo intellekta. no 2(37), 2025: 79–90.

RESUME

K.A. Nikitenko, A.V. Zviagintseva

Assessment of semantic similarity of sentences using probabilistic measures based on embeddings

Modern semantic analysis in natural language processing widely relies on vector representations of words and attention-based neural architectures. However, most existing approaches describe semantic relations geometrically and do not provide explicit probabilistic interpretation of meaning. The purpose of this paper is to develop a method for evaluating semantic similarity of sentences based on the comparison of independent and context-dependent probabilistic events derived from word embeddings and attention weights.

The study introduces a probabilistic method in which words are treated as random events, and sentences are represented as distributions over independent and dependent probabilities. Probabilistic embeddings are used to model lexical meaning, while an attention-based mechanism is applied to estimate contextual dependencies between words. A comparative analysis of these probability spaces makes it possible to quantify semantic coherence at the sentence level.

The proposed method allows formalizing semantic relations between words and reveals clear distinctions between random word co-occurrence and contextually conditioned connections. The analysis shows that sentences with stronger internal semantic structure demonstrate significantly higher coherence within the probabilistic model, confirming the adequacy of the approach for interpreting meaning in text.

The developed method provides an interpretable probabilistic alternative to purely geometric similarity metrics and can be applied in semantic search, text clustering, and evaluation of generated texts. Future work includes extending the method to larger corpora and integrating contextualized embeddings from transformer models to refine probabilistic estimation of semantic relations.

РЕЗЮМЕ

К.А. Никитенко, А.В. Звягинцева
Оценка семантической схожести предложений с использованием вероятностных мер эмбеддингов

Современные методы обработки естественного языка активно используют распределённые представления слов и механизмы внимания, однако большинство подходов описывает смысловые связи только геометрически, без явной вероятностной интерпретации. Актуальной задачей является разработка методов, позволяющих формализовать степень смысловой связанности текста и оценивать схожесть предложений с опорой на статистические свойства языковых моделей. Цель работы – предложить вероятностный подход к оценке семантической схожести, основанный на сопоставлении независимых и контекстно-зависимых событий, соответствующих словам в предложении.

В работе предложен вероятностный метод оценки семантической схожести предложений, основанный на сопоставлении независимых и зависимых событий, соответствующих словам в предложении. Метод объединяет вероятностные распределения плотностей эмбеддингов и матрицы внимания, что позволяет количественно описать степень смысловой связанности текста.

Результаты анализа показывают, что предложенный метод дает возможность различать случайные сочетания слов и контекстно обусловленные связи, что делает возможной формализацию семантической структуры текста в вероятностных терминах. Метод демонстрирует устойчивое поведение при анализе корпуса и корректно выделяет предложения с выраженной смысловой связностью.

Предложенный метод позволяет формализовать понятие семантической когерентности текста в вероятностных терминах и применять его в задачах семантического поиска, кластеризации предложений, оценки связности автоматически генерируемых текстов и анализа интерпретируемости нейросетевых моделей.

Никитенко Кирилл Андреевич – аспирант кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», ул. Университетская, 24, Донецк, 283001, Россия, n1kitenkok@yandex.ru.
Область научных интересов: компьютерное зрение, машинное обучение, нейронные сети. Число научных публикаций – 5.

Звягинцева Анна Викторовна – д.т.н., доцент, профессор кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», ул. Университетская, 24, Донецк, 283001, Россия, zviagintsevaav@gmail.com. *Область научных интересов*: системный анализ, событийная и комплексная оценка; безопасность и управление социально-экономическими и техногенными системами; информационно-аналитические системы; обработка и анализ данных. Число научных публикаций – более 150.

Статья поступила в редакцию 29.08.2025