УДК 004.02

# СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ СБОРА И ИСПОЛЬЗОВАНИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ МОДЕЛЕЙ В МЕДИЦИНЕ

**Н. Г. Воробьев**, аспирант Московского политехнического университета, г. Москва, Россия; orcid.org/0000-0001-5995-7526, e-mail: nickikta@yandex.ru

***Цель работы*** *– оценить методы автономного сбора и использования локальных контекстов при выборе варианта ответа интеллектуальной системой в процессе диалогового взаимодействия.*

*В работе для анализа использовались тексты технической документации и научных публикаций, соответствующих предметной области «медицина». Рассмотрены современные методы сбора локальных контекстов, а также способы формализации текстовых данных для последующей обработки программными средствами.*

*Предложен новый метод графового моделирования логических связей естественно-языковых данных. Данная формализация позволяет сохранить логические связи элементов с окружающим текстом, что делает ее полезной как для лингвистических исследований, направленных на изучение особенностей употребления слов в различных сферах деятельности, так и для оптимизации алгоритмов выбора ответов в диалоговых системах. Проведено сравнение эффективности различных подходов к моделированию локальных контекстов.*

***Keywords:*** *искусственный интеллект, локальный контекст, диалоговое взаимодействие, тезаурус, чат-бот.*

**DOI:** 10.21667/1995-4565-2025-93-200-212

## Введение

Интеграция информационных технологий в современную медицину становится все более распространенной, при этом аналитические и прогностические системы играют значительную роль в повышении качества диагностики и лечения. Важной категорией таких технологий являются экспертные системы, которые предназначены для предоставления рекомендаций на основе заранее формализованных знаний.

Важная проблема проектирования диалоговых систем связана со сбором и обработкой данных для разработки и обучения языковых моделей. Предметная область «медицина» характеризуется большим количеством контекстуальных синонимов и сложной терминологией, что затрудняет создание моделей без привлечения экспертов. Кроме того, доступ разработчиков к полным и высококачественным данным исследований ограничен из соображений конфиденциальности и характера медицинской информации.

В данной работе проведен анализ существующих методов сбора и оценки качества локальных контекстов. Под локальным контекстом в данном случае понимается совокупность слов и их связанных семантических полей, необходимых для наиболее полного понимания включающих их высказываний [1]. Наиболее распространенным способом представления такой информации являются тезаурусы. Технология их создания подробно рассматривается в [2]. Существует множество методов анализа текстов для извлечения необходимой информации. Один из базовых подходов подробно рассматривается в [3].

Учитывая трудности с доступом к закрытым исследовательским данным, в этом исследовании используются материалы из научных работ с открытым доступом в онлайн библиотеках. В рамках исследования эффективность естественно-языковых моделей оценивается по качеству категоризации, производимой с их помощью. В частности, это включает в себя при-

своение описания на естественном языке, связанного с конкретным заболеванием, соответствующему контекстуальному полю.

Существует множество методов анализа текстов для извлечения необходимой информации [4]. Далее в работе детально рассматриваются основные алгоритмы сбора локальных контекстов, способы кодирования полученных данных и метрики для проверки эффективности созданного тезауруса.

## Материалы и методы

Основная цель исследования состоит в разработке подхода к моделированию локального контекста, который позволит системе принятия решений динамически анализировать и интерпретировать контекст для повышения релевантности и точности ответов.

Был проведен углубленный анализ существующих алгоритмов обработки текстовой информации. Особое внимание было уделено методам, поддерживающим контекстуальное моделирование, поскольку они имеют решающее значение для понимания тонкостей естественного языка.

Изучалась возможность применения нового графового метода моделирования. Этот этап включал не только теоретическое исследование, но и практические эксперименты для оценки применимости и эффективности предложенного метода. Значительное внимание было уделено установлению сравнительных критериев для систематической оценки эффективности различных подходов.

Объем исследования был ограничен тремя основными методами моделирования локальных контекстов. Эти методы были выбраны для предоставления репрезентативной выборки существующих подходов. Для теоретической части были выбраны научные источники не позднее 2017 года публикации.

На экспериментальной стадии исследования использовались текстовые данные из предметной области «медицина», причем все исходные тексты написаны на английском языке. Этот предметно-ориентированный фокус был выбран для решения прикладных задач. Модели, созданные на этом этапе, были подвергнуты оценке эффективности.

## Результаты

Для понимания того, что такое локальный контекст рассмотрим принцип работы чатбота. Он функционирует как стандартный виртуальный помощник: подключается к диалоговой платформе, разработчик задает перечень запросов и соответствующих им ответов, вносит их в базу данных и активирует систему [5]. В процессе взаимодействия с пользователями бот отвечает по заранее заданным сценариям, одновременно фиксируя слова и фразы, сопровождающие вопрос и ответ. Эти элементы формируют локальный контекст ресурса.

### Статистический анализ

Статистический анализ представляет собой ключевой метод, применяемый для сбора локального контекста [6]. Входящий пользовательский запрос очищается от незначащих слов, после чего сравнивается со всеми вопросами, хранящимися в базе данных. Если уровень совпадения превышает установленный порог идентичности, система считает вопрос распознанным и возвращает идентификатор строки базы данных, содержащей соответствующий ответ.

Если совпадение достигает лишь порогового значения потенциальной идентичности, пользователю предлагаются дополнительные уточняющие вопросы, на основе которых подбирается наиболее релевантный ответ. При этом все слова, встречающиеся в пользовательском запросе и отличающиеся от слов в распознанном вопросе, фиксируются в отдельной таблице базы данных как контекстные синонимы. Эти данные сохраняются в формате «слово» – «синоним» – «частота» [7]. При первом добавлении пары слов частотный показатель устанавливается в единицу и увеличивается на единицу с каждым последующим обнаруже-

нием. В дальнейшем эта информация используется в статистическом анализе для повышения эффективности системы.

При обработке входящего запроса, помимо исходного пользовательского ввода, сохраняется массив альтернативных формулировок, в которых слова из таблицы синонимов заменяются их эквивалентами. Благодаря этому анализ охватывает не только сам запрос, но и его возможные смысловые вариации, что значительно повышает точность распознавания. Таблица синонимов представляет собой локальный контекст, актуальный для текущего взаимодействия с системой [8].

Алгоритм работы:

- система сравнивает заданный вопрос с каноническими вариантами;
- при достижении порога идентичности задается уточняющий вопрос;
- сохраняет альтернативные слова и формирует локальный контекст;
- использует локальный контекст для выбора ответа на последующих этапах.

В этом случае системой будет легко управлять, но правильная настройка параметров очень затруднительна и требует экспертного вмешательства.

**Тезаурусный поиск**

Второй метод – тезаурусный поиск [9]. Для его использования помимо данных, содержащихся в базе данных, необходим сам информационно-поисковый тезаурус. Это файл, который включает информацию в формате «слово» – «лемма» – «частота» и другие поля. В данном исследовании был использован тезаурус, оформленный в XML.

Модуль разделяет ввод пользователя, очищенный от лишних слов, на двусловия и обрабатывает их. Каждой паре слов соответствует своя лемма [10]; если для пары лемма не найдена, то она ищется для каждого слова отдельно. Массив лемм очищается от дублирующихся элементов и сохраняется как обработанный пользовательский запрос. Если уровень совпадения превышает установленный порог, вопрос считается распознанным, и возвращается ответ.

Алгоритм работы:

- нормализует запрос;
- разбивает запрос на двусловия;
- извлекает леммы из двусловий;
- сравнивает леммы с леммами, соответствующими известным запросам.

В то же время система будет сразу готова к работе и покажет стабильные результаты распознавания, но процесс работы значительно замедлится из-за множества обращений к базе данных. Необходимость разработки собственного тезауруса или получения готового также усложняет применение метода.

**Интеллектуальный поиск**

Интеллектуальный поиск осуществляется с использованием нейронной сети [11]. Она обучается на примерах корректных и некорректных запросов. Ввод пользователя проходит очистку от незначимых слов формализуется с помощью обученной модели. Затем происходит анализ и выбор ответа ответ [12]. Такая система будет работать эффективно, минимизируя количество запросов к базе данных, однако обучение нейронной сети может занять продолжительное время. Кроме того, для эффективного обучения требуется значительный объем размеченных данных, что влечет за собой увеличение сложности или требовательности к экспертному подходу.

**Методы формализации естественного языка**

Теперь рассмотрим методы обработки слов. В связи с ограничениями нейронных сетей, которые принимают входные данные только в формализованном виде [13], необходимо применять методику приведения слов к числовому эквиваленту с сохранением обратной смысловой связи.

Первый из таких методов – это «мешок слов». Рассмотрим подробнее, как он работает. Каждое слово кодируется по частоте его появления в обучающей выборке. Идентификаторы, сохраненные таким способом, имеют обратную логическую связь с соответствующими словами, но не содержат данных о грамматике или порядке слов [14]. Проблема этого подхода заключается в том, что для его правильного функционирования пользователь должен вводить слова без ошибок, потому что слово с ошибкой будет интерпретироваться как отдельное.

Второй метод обработки слов – это использование векторов слов. Слово представляется вектором в трехмерном пространстве, где расстояние между точками определяется контекстуальной близостью по набору параметров. Расстояние между векторами отражает контекстную близость: слова, которые часто встречаются рядом в одном контексте, будут иметь схожие векторы внутри обученной модели [15].

**Расчет эффективности тезауруса**

Эффективность тезауруса определяется его информационной полнотой [16]. Для оценки этой полноты используется индекс энтропии, который рассчитывается на основе элементов, содержащихся в словаре. В дальнейшем количество слов в тезаурусе будет обозначаться символом $n$. В данном случае каждое слово является существительным и вероятность перехода существует лишь между ними. Если вероятность выбора любого слова одинакова и случайна, то расчет энтропии будет представлен формулой (1):

$$H\left(\frac{1}{n}\right) = -\frac{1}{n}\ln\left(\frac{1}{n}\right) - \left(1-\frac{1}{n}\right)\ln\left(1-\frac{1}{n}\right), \quad (1)$$

где $H$ – энтропия отдельного элемента, а $n$ – количество слов, содержащихся в тезаурусе.

Как следует из формулы, результат расчета зависит не только от числа элементов в словаре, но и от выбранного основания логарифма [17]. Для упрощения расчетов можно применить натуральный логарифм. Для более наглядного представления следует вычислить коэффициент энтропии и коэффициент вариации тезауруса, которые выражаются формулах (2) и (3):

$$A\left(\frac{1}{n}\right) = \frac{1}{n}\left(\ln\left(\frac{1}{n}\right)\right)^2 - \left(1-\frac{1}{n}\right)\left(\ln\left(1-\frac{1}{n}\right)\right)^2, \quad (2)$$

$$\sigma(n) = \sqrt{n\left[A\left(\frac{1}{n}\right) - H\left(\frac{1}{n}\right)\right]^2}, \quad (3)$$

где $A$ – это коэффициент энтропии, $n$ – количество слов в тезаурусе, σ – коэффициент вариации тезауруса, а $H$ – энтропия отдельного элемента.

Полученные уравнения могут быть использованы для построения графиков этих функций. При этом можно заметить, что параметр мощности растет пропорционально числу слов в тезаурусе, а график коэффициента вариации стремится к единице.

Следующей характеристикой является точность, которая отражает соотношение правильно выбранных ключевых слов и общего числа ключевых слов. Полнота показывает отношение выбранных ключевых слов к числу всех возможных совпадений. Из этих двух характеристик вычисляется гармоническое среднее, которое показывает баланс между ними. Качество собранных данных также зависит от количества связей в тезаурусе. Чем больше семантических связей между словами и выше уровень связности, тем более эффективно можно применять собранные данные. Для оценки этого параметра используется граф, отображающий отношение множества вершин, представляющих термины тезауруса и множеству связей между этими терминами.

Количество терминов и их взаимосвязи являются важнейшими метриками для оценки тезауруса [18]. Количество ключевых слов должно быть достаточно большим, а количество связей должно быть сопоставимо с количеством терминов, чтобы избежать как избыточной

нагрузки, так и недостаточной связности слов. В случае нарушения этих условий тезаурус становится менее эффективным для использования.

### Сравнение эффективности тезаурусов

Для сравнения были выбраны несколько систем сбора тезаурусов. Первой является Thesaurus.com, дочерний проект Dictionary.com. Он является основным онлайн-ресурсом для поиска синонимов в Интернете [19]. Второй тезаурус называется RuWordNet. Он построен на автоматической переработке тезауруса RuThes в формат WordNet [20]. Этот ресурс содержит синонимы для существительных, глаголов и прилагательных и включает более 100 000 ключевых слов и фраз на русском языке. В нем также выделяются различные типы связей между словами. Третий тезаурус, ABCThesaurus, является многоязычным и насчитывает более 14 000 синонимов. Он используется как вспомогательное средство для писателей. Последний тезаурус, который участвует в сравнении, был составлен экспертами – выпускниками медицинских вузов специально для этой работы.

Для сбора данных, используемых в сравнении тезаурусов, был создан массив ключевых слов на основе анализа научных работ по фармакологии. Тексты трех работ были объединены, очищены от служебных слов и отсортированы по частоте употребления уникальных слов. В результате были выбраны 150 самых часто встречающихся слов. Эти слова использовались как ключевые при сборе тезаурусов с помощью упомянутых выше ресурсов. Четвертый тезаурус был составлен с использованием экспертного подхода.

Итоговый набор характеристик позволяет всесторонне оценить качество высокоспециализированного тезауруса.

На основе полученных данных можно сделать несколько выводов, представленных в виде графиков. В частности, полнота тезауруса оценивается как соотношение объема собранной информации. Здесь лидирует thesaurus.com благодаря большому количеству обновленных слов. Коэффициент полноты (рисунок 1) демонстрирует оптимальный состав массива слов, и идеальное значение этого критерия должно стремиться к единице, хотя ни один из анализируемых тезаурусов не достиг этого уровня.


**Рисунок 1 – Полнота тезаурусов**
**Figure 1 – Thesauri completeness**

Также можно отметить, что эксперты собрали почти идентичное количество слов, что и ABCThesaurus, и RuWordNet. Система thesaurus.com, благодаря регулярному обновлению своей базы данных и гибким настройкам, демонстрирует уровень связности, который близок к экспертному, как показано на графике. Это указывает на склонность системы выявлять логические связи в тексте, а не искать синонимы вне его в заранее подготовленной базе данных. Однако такой результат также обеспечивается большим количеством слов, которые си-

стема добавляет в свой словарь, что видно из коэффициента отдаления (рисунок 2), который оказался достаточно низким. Этот коэффициент значительно выше для тезауруса, составленного экспертами, что указывает на взаимосвязь между методом создания словаря и «человечностью» слов, которые в него включены. При этом стремление этого параметра к единице свидетельствует о более «ассоциативных» связях в тезаурусе, в то время как большое расстояние указывает на преобладание «синонимных» связей. Ассоциативные отношения предполагают наличие различных слов, между которыми при анализе определенного источника информации образуются регулярные связи. Синонимы же – это разные слова с похожими значениями [21].

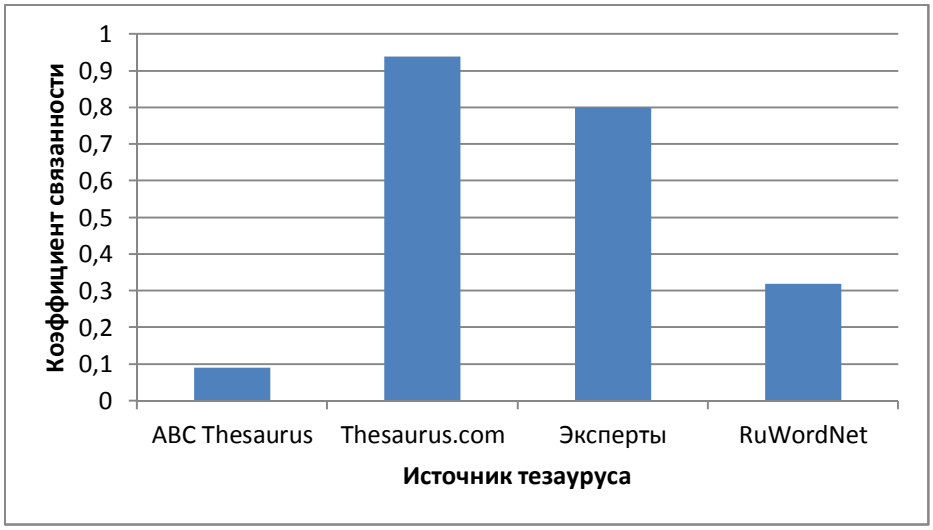


**Рисунок 2 – Связанность терминов**
**Figure 2 – Terms relatedness**

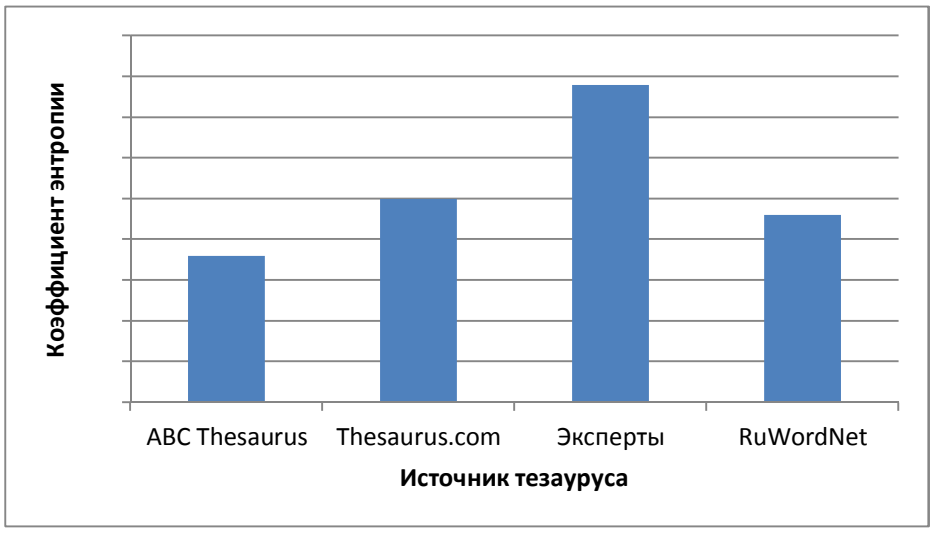


**Рисунок 3 – Энтропия терминов**
**Figure 3 – Terms entropy**

Данные, полученные экспертами, обладают наибольшей упорядоченностью, что можно увидеть на рисунке 3. Тем не менее, несмотря на различия в подходах к созданию тезаурусов и ориентированности систем, автоматические методы сбора демонстрируют почти одинаковые результаты по этому показателю. Коэффициент вариации (рисунок 4) для thesaurus.com значительно превосходит всех конкурентов, что объясняется его значительным преимуществом в количестве выбранных слов.


**Рисунок 4 – Коэффициент вариаций**
**Figure 4 – Variations coefficient**

Теперь, когда методы формализации были обозначены, необходимо перейти к способам моделирования и представления локальных контекстов.

### Моделирование локальных контекстов предметных областей

Для демонстрации и сравнения моделей будет применяться измерение эффективности рубрицирования естественно-языковых текстов. Таким образом, тема одного или нескольких текстов будет обозначать категорию запросов, а отдельные документы или фрагменты текста из категории – подкатегорию. Эффективностью рубрицирования будет считаться точность отнесения подкатегорий к категории автоматической системой, основанной на рассматриваемой модели.

Подкатегории предметных областей задаются естественно-языковыми описаниями, содержащимися в научных статьях, таким образом можно сделать вывод о том, что они характеризуются следующей спецификой. Во-первых, они более формальные, чем общеупотребительная прямая речь, но менее формальные, чем академические учебные тексты из учебников. Помимо этого статьи, предоставленные для обработки, характеризуют одну глобальную категорию заболеваний и отличаются описаниями особенностей протекания болезни. Соответственно, контекстные поля работ достаточно близки, как и используемая лексика. Это значительно усложняет создание подходящей модели.

Ввиду этого было решено сравнить эффективность нескольких различных подходов к моделированию контекстуальных полей. А именно: отдельные модели, характеризующие подкатегории и метод выбора, основанный на сравнении весов, большая модель, содержащая все представление предметной области, нейронная сеть, обученная на ней, и новый метод моделирования – отдельные графовые модели, описывающие подкатегории и метод выбора, основанные на сравнении близости направленных графов.

В начале работы были выбраны четыре эталонные научные статьи, характеризующие подкатегории. Данные работы предложили эксперты из числа представителей медицинского вуза. Объем информации, содержащейся в четырех исходных статьях, оказался недостаточным для полноценного обучения. В связи с этим было принято решение расширить объем данных, погрузившись на ссылочный уровень. Было выдвинуто предположение, что статьи, упомянутые в списке источников каждой из исследуемых статей, имеют непосредственное отношение к их тематике и могут быть использованы для дополнения выборки.

Таким образом, для каждой из четырех исходных подкатегорий был сформирован массив из 100 статей, указанных в их библиографических ссылках. В тех случаях, когда одна статья присутствовала в списке источников нескольких подкатегорий, она исключалась из выборки.

Далее применялся метод углубления на следующий ссылочный уровень, чтобы достичь обозначенного количества текстов.

Таким образом, был создан итоговый датасет, включающий 400 научных статей, распределенных по четырем категориям протекания заболевания. Этот подход позволил обеспечить достаточный объем данных для обучения моделей.

**Отдельные векторные модели**

Первый подход к решению задачи предполагает использование отдельных векторных моделей для каждой из подкатегорий. В данном методе локальные контексты описываются отдельными моделями Word2Vec, которые создаются с небольшой размерностью вектора для оптимизации вычислений. Каждая такая модель обучается на своем собственном датасете, соответствующем конкретной подкатегории. При этом можно использовать и другие виды моделей [22], специфичных для подходящей задачи.

В результате обучения каждая модель представляет собой набор нормализованных слов и их векторных представлений. Все четыре модели и их идентификаторы сохраняются в базе данных, что позволяет эффективно управлять данными и проводить анализ поступающих текстов.

Когда система получает запрос, например текст новой статьи, для определения ее категории выполняется следующий алгоритм.

1. Запрос анализируется с использованием каждой из четырех сохраненных моделей.
2. Для текста рассчитываются векторы по каждой модели.
3. Рассчитанные векторы сравниваются между собой.
4. Модель, которая выдала вектор с наибольшим значением, считается наиболее релевантной, а ее подкатегория – наиболее близкой к контексту статьи.

Этот метод демонстрирует высокую эффективность при работе с небольшими текстами, которые легко классифицируются по своему контексту. В данном подходе каждая модель фактически выступает в роли локального контекста, отражающего узкое поле изучаемой предметной области. Такой подход упрощает категоризацию и делает процесс анализа более точным и структурированным.

**Общая векторная модель**

Второй подход основан на использовании одной общей модели, созданной на всем датасете [23], и нейросети, которая обучается на этой модели для распознавания подкатегорий. В данном случае векторная модель создается с помощью Word2Vec с большей размерностью, что позволяет учесть более широкий контекст и охватить всю предметную область целиком.

Процесс обучения модели включает два этапа. Сначала она предобучается на общем датасете, содержащем текстовые данные общего назначения, чтобы сформировать базовые представления. Затем выполняется дообучение на основном датасете, который ранее был собран из ссылок. Это позволяет модели адаптироваться к специфике предметной области. В отличие от первой методики, эта модель работает не со словами, а с токенами, что делает ее более универсальной. Модель сохраняется как отдельный компонент системы и не идентифицируется внутри проекта, так как она является уникальной для данной задачи.

Для распознавания поступающих запросов, таких как текст новой научной статьи, требуется обучение нейронной сети. Обучение продолжается до тех пор, пока сеть не достигнет высокой точности категоризации текстов из обучающего набора данных. После завершения обучения сеть становится способной определять подкатегорию любого нового текста.

Процесс обработки запроса включает расчет сложного вектора для нового текста с использованием общей модели. Этот вектор поступает в нейросеть, которая определяет его подкатегорию.

Этот подход демонстрирует свою эффективность при работе с большими текстами, где требуется учитывать сложные и трудно разделяемые контексты. Вместо конкретных сущностей модель формирует массив специфичных векторов, который служит контекстным обра-

зом для каждой подкатегории. Нейронная сеть обучается именно на классификации этих контекстных векторов, что позволяет достичь высокой точности даже для сложных запросов.

**Графовая модель**

В рамках данной работы также был разработан альтернативный метод моделирования локальных контекстов с помощью графов. В этой модели вершинами выступают слова, а веса ребер характеризуют связанность между ними. Связанность определяется как мера отдаленности слов друг от друга.

Существует два варианта создания графовой модели [24]. Первый вариант предполагает создание полной модели, включающей все слова из датасета, что обеспечивает максимальную точность. Второй вариант – упрощенная модель, которая состоит только из лемм слов. Упрощенный подход снижает размерность модели, но при этом несколько теряет в точности.

Обучение графовой модели сводится к заполнению графа, где веса ребер устанавливаются на основании связей между словами [25]. Для реализации системы принятия решений создаются отдельные графовые модели для каждой подкатегории на основе соответствующих датасетов. Когда поступает пользовательский запрос (например, текст новой статьи), он преобразуется в небольшой граф. Этот граф затем сравнивается с графами подкатегорий методом расчета близости направленных графов – формула (4), где d(G,F) – нормализованная величина близости графов, mcs – максимально общий подграф контекстуальных графов, $min_{i=1,\ldots,k}$ – минимальное число общих вершин, g и f – сравниваемые подграфы.

$$d(G,F) = 1 - min_{i=1,\ldots,k}\left(\frac{\left|mcs(g_{\min(i,m),f_1})\right|}{i}\right). \quad (4)$$

В таком случае формула модели определяется как: G = (N, C, W), где N – вершины графа (языковые единицы, составляющие грамматику); C – дуги, объединяющие вершины; W – веса дуг (вероятность перехода между вершинами).

При необходимости, для повышения точности анализа, можно углубиться в связанность на следующих уровнях графа, исследуя дополнительные связи между словами.

Основным недостатком данного метода является высокая требовательность к объему памяти для хранения графовых моделей, а также значительное время, требуемое для обработки запросов. Тем не менее, этот подход обеспечивает самую высокую точность благодаря использованию графовых метрик.

Контекстное поле подкатегории в данном методе описывается взвешенным графом, что позволяет очень точно относить небольшие тексты к одной из подкатегорий. Такой подход особенно эффективен, когда требуется максимально точная классификация.

**Сравнение эффективности моделей**

В ходе работы были получены следующие результаты.

1. При использовании нескольких векторных моделей запросы длиной до 700 символов распознавались с эффективностью 61 %. Этот метод показал свою пригодность для предварительной рубрикации текстов, когда требуется быстрая, но не максимально точная классификация.

2. Метод с одной векторной моделью и нейронной сетью продемонстрировал высокую эффективность. Для предобучения модели использовался датасет, составленный из учебников по медицине, а дообучение проводилось на 400 статьях, собранных по предметной области. Система успешно обрабатывала запросы длиной до 10 000 символов, достигая корректной категоризации в 76 % случаев для сложных научных текстов. Этот подход показал себя как практичный инструмент для работы с длинными текстами и может быть эффективно применим в реальных условиях.

3. Графовая модель обучалась на кратких текстах, извлеченных из аннотаций научных статей. Запросы длиной до 50 символов были распознаны с точностью 100 %. Однако из-за

длительного времени анализа этот метод оказался менее подходящим для практического использования, несмотря на его высокую точность.

4. Таким образом, каждый из методов продемонстрировал свои сильные и слабые стороны, что позволяет выбирать подход в зависимости от требований к скорости, длине текста и точности классификации.

## Заключение

В результате работы поставленная задача была успешно решена. Также были сделаны несколько важных выводов.

1. Наиболее простой и эффективный способ обучения моделей – это прямое обучение Word2Vec на текстах, относящихся к предметной области. Такой подход минимизирует сложность подготовки данных и обеспечивает базовую функциональность.

2. Создание более сложных датасетов требует предварительной разметки [26], что увеличивает трудозатраты, но позволяет существенно повысить точность распознавания запросов.

3. Для достижения более высокой эффективности можно использовать отдельные модели в качестве контекстных полей терминов. Эти модели будут служить основой для дальнейшего анализа, где запросы будут оцениваться с помощью графового метода. В этом случае подкатегорию можно представить исключительно через набор ключевых терминов, что упрощает и структурирует процесс классификации.

4. В графовой реализации необходимо перейти к системе, при которой вершинами графа выступают отдельные векторные модели, а ребра отражают взаимосвязи между ними. Это позволит не только учитывать структуру терминологических связей, но и проводить более точную категоризацию запросов. Это позволит улучшить эффективность существующих методов категоризации текстов [27].

Собранные в графовую модель данные имеют структуру, которая позволяет каждому ключевому слову сопоставить локальный контекст либо сопоставлять контекстуальный граф целой категории. Во втором случае подкатегории будут характеризоваться областями этого графа. Этот контекст представлен в формате «ключ-значение», где ключом является само слово, а значением – вероятность перехода к другим словам или элементам, с учетом анализируемого контекста. Такая структура обеспечивает возможность гибкого и точного представления взаимосвязей между словами на основе их окружения.

Таким образом, разработанная графовая модель обладает потенциалом для успешного использования в задачах анализа данных и обработки естественного языка.

Данные выводы подчеркивают значимость выбора подходящего метода в зависимости от задач и позволяют адаптировать систему под конкретные требования.

*Настоящее исследование проведено при финансовой поддержке Московского Политехнического Университета в рамках гранта имени В.Е Фортова*.

## Библиографический список

1. Philippe SchlenkerInstitut, Local Contexts, Semantics & Pragmatics, 2009. Vol. 2. PP. 1-78 (2009).

2. **Sosnina E.P.**, Designing an educational thesaurus as a component of a translation course in the oil and gas industry, Proceedings of the Samara Scientific Center of Russian Academic Sciences, 22, 2020. PP. 38-42.

3. **Tsitulsky A.M., Ivannikov A.V., Rogov I.S.**, Intellectual text analysis. StudNet, 2019. Vol. 6. PP. 476-482.

4. **Loukachevitch N., Alekseev A.** Use of neighbor sentence co-occurrence to improve word semantic similarity detection. 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT). 10.1109/FRUCT. 2016. 7584768

5. **Sobirov B.I., Nurmetova B.B.**, Use bot chats with artificial intelligence in the sphere of telecommunications to reduce the queue to operators, Academic Research in Educational Sciences. 2022. PP. 312-315.

6. **Hernández-Sánchez J., Grunchec J.-A., Knott S**. A web application to perform linkage disequilibrium and linkage analyzes on a computational grid, BIOINFORMATICS. 2009. PP. 1377-1383.

7. **Keith E., Kent A.** Representing Thoughts, Words, and Things in the UMLS, Journal of the American Medical Informatics Association. 1998. Vol. 5(5). PP. 421-431.

8. **Schulz E.B**, **Barrett J.W**., MB, BCHIR, MSC, Colin price, MPHIL, FRCS, Read Code Quality Assurance: From Simple Syntax to Semantic Stability (Journal of the American Medical Informatics Association. 1998. Vol. 5(4). PP. 337-346.

9. **Amberger J.S., Bocchini C.A., Schiettecatte F., Alan F. Scott A.F., Hamosh A**. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders, Nucleic Acids Research. 2015. Vol. 43. PP. 789-798.

10. **Bańko M.** Lemmatization Algorithms for Dictionary Users. A Case Study, International Journal of Lexicography. 1992. Vol. 5(3). PP. 199-220.

11. **Yao Z., Zhang W., Song P., Hu Y., Liu J.** DeepFormer: a hybrid network based on convolutional neural network and flow-attention mechanism for identifying the function of DNA sequences, Briefings in Bioinformatics. 2023. Vol. 24(2).

12. **Sosnina E.P.** Designing an educational thesaurus as a component of a translation course in the oil and gas industry, Proceedings of the Samara Scientific Center of Russian Academic Sciences. 2020. Vol. 22. PP. 38-42.

13. Darja Fišer and Nikola Ljubešić, Distributional modeling for semantic shift detection, International Journal of Lexicography. 2019. Vol. 32(2). PP. 163-183.

14. **Wool J.R.** Does the Cryptographic Hashing of Passwords Qualify for Statutory Breach Notification Safe Harbor? Journal of Law & Cyber Warfare. 2018. Vol. 6(2). PP. 56-89.

15. **Zhu F., Fellbaum C**. Quantifying Fixedness and Compositionality in Chinese Idioms (International Journal of Lexicography. 2015. Vol. 28(3). PP. 338-350.

16. **Beckwith R., Miller G.** Implementing a Lexical Network, International Journal of Lexicography. Vol. 3(4). P P . 302-312.

17. **Tanyimboh T.T., Templeman A.B**. Calculating Maximum Entropy Flows in Networks, The Journal of the Operational Research Society, New Research Directions. 1993. Vol. 44(4). PP. 383-396.

18. **Fontenelle T.** WordNet, FrameNet and Other Semantic Networks in the International Journal of Lexicography – The Net Result? International Journal of Lexicography. 2012. Vol. 25(4). PP. 437-499.

19. **Gilquin G., Laporte S.** The Use of Online Writing Tools by Learners of English: Evidence From a Process Corpus, International Journal of Lexicography. 2021. Vol. 34(4). PP. 472-492.

20. **Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J.** Introduction to WordNet: An Online Lexical Database, International Journal of Lexicography. 1990. Vol. 3(4). PP. 235-244.

21. **Foxley E., Gwei G.M.** Synonymy and Contextual Disambiguation of Words, International Journal of Lexicography. 1989. Vol. 2(2). PP. 111-134.

22. **Suyatinov S.I., Buldakova T.I., Vishnevskaya J.A**. Identification of Situations Based on Synergetic Model. 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA). 2021.10.1109/SUMMA53307.2021.9632207.

23. **Dudnikov S., Miheev P., Grinkina T.** Evaluating of Word Embeddings Hyper-parameters of the Master Data in Russian-Language Information Systems. In: Hu, Z., Petoukhov, S., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics. CSDEIS 2019. Advances in Intelligent Systems and Computing 2020. Vol. 1127. Springer, Cham. https://doi.org/10.1007/978-3-030-39216-1_7.

24. **Belyanova M., Chernobrovkin S., Latkin I., Gapanyuk Y.** Metagraph Based Approach for Neural Text Question Generation. In: van der Aalst, W.M.P., et al. Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. 2021. Vol. 12602. Springer, Cham. https://doi.org/10.1007/978-3-030-72610-2_6.

25. **Kanev A., Terekhov V., Chernenky V., Proletarsky A.** Metagraph Knowledge Base and Natural Language Processing Pipeline for Event Extraction and Time Concept Analysis. 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). 2021. 10.1109/ElConRus51938.2021.9396541.

26. **Savchenko M., Tynchenko V.** Unsupervised Production Machinery Data Labeling Method Based on Natural Language Processing. International Russian Smart Industry Conference (SmartIndustryCon). 2024. 10.1109/SmartIndustryCon61328.2024.10515763.

27. **Rogov A.A., Loukachevitch N.V**. Evaluating the Performance of Interpretability Methods in Text Categorization Task. Lobachevskii J Math 45. 2024. PP. 1234-1245.

UDC 004.02

# EFFICIENCY COMPARISON OF METHODS TO COLLECT AND USE NATURAL LANGUAGE MODELS IN MEDICINE

**N. G. Vorobyov**, post graduate student, Moscow Polytechnic University, Moscow, Russia; orcid.org/0000-0001-5995-7526, e-mail: nickikta@yandex.ru

*The aim of this study is to evaluate methods for autonomous collection and usage of local contexts when selecting a response option by intelligent system during dialog interaction.*

*The analysis uses texts from technical documentation as well as scientific publications related to medical subject area. Modern methods to collect local contexts, as well as methods to formalize text data for subsequent processing by software are considered.*

*A new method for graph modeling of logical relationships in natural language data is proposed. This formalization preserves logical connections between elements and surrounding text making it useful both for linguistic research aimed at studying the specifics of word usage in various fields and for optimizing response selection algorithms in dialogue systems. The efficiency of various approaches to modeling local contexts has been compared.*

***Keywords:*** *Artificial Intelligence, Local context, Dialog interaction, Thesaurus, Chatbot.*

**DOI:** 10.21667/1995-4565-2025-93-200-212

**References**

1. Philippe Schlenker Institut, *Local Contexts, Semantics & Pragmatics*, 2009, vol. 2, pp. 1-78.

2. **Sosnina E.P.**, Designing an educational thesaurus as a component of a translation course in the oil and gas industry, *Proceedings of the Samara Scientific Center of Russian Academic Science*, 2020, vol. 22, pp. 38-42.

3. **Tsitulsky A.M., Ivannikov A.V., Rogov I.S.** Intellectual text analysis. student. 2019, vol. 6, pp. 476-482.

4. **Loukachevitch N., Alekseev A.** Use of neighbor sentence co-occurrence to improve word semantic similarity detection. *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*. 10.1109/FRUCT. 2016. 7584768

5. **Sobirov B.I., Nurmetova B.B.** Use bot chats with artificial intelligence in the sphere of telecommunications to reduce the queue to operators, *Academic Research in Educational Sciences*. 2022, pp. 312-315.

6. **Hernández-Sánchez J., Grunchec J.-A., Knott S.** A web application to perform linkage disequilibrium and linkage analyzes on a computational grid, B*IOINFORMATICS*. 2009, pp. 1377-1383.

7. **Keith E., Kent A.** Representing Thoughts, Words, and Things in the UMLS, *Journal of the American Medical Informatics Associatio*n. 1998, vol. 5(5), pp. 421-431.

8. **Schulz E.B, Barrett J.W.**, MB, BCHIR, MSC, Colin price, MPHIL, FRCS, Read Code Quality Assurance: From Simple Syntax to Semantic Stability. *Journal of the American Medical Informatics Association.* 1998, vol. 5(4), pp. 337-346.

9. **Amberger J.S., Bocchini C.A., Schiettecatte F., Scott A.F.,Hamosh A.** OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders, Nucleic Acids Research. 2015, vol. 43, pp. 789-798.

10. **Bańko M.** Lemmatization Algorithms for Dictionary Users. A Case Study*, International Journal of Lexicography*. 1992, vol. 5(3), pp. 199-220.

11. **Yao Z., Zhang W., Song P., Hu Y., Liu J.** DeepFormer: a hybrid network based on convolutional neural network and flow-attention mechanism for identifying the function of DNA sequences, *Briefings in Bioinformatics*. 2023, vol. 24(2).

12. **Sosnina E.P.** Designing an educational thesaurus as a component of a translation course in the oil and gas industry. *Proceedings of the Samara Scientific Center of Russian Academic Sciences*. 2020, vol. 22, pp. 38-42.

13. **Fišer D., Jubešić N.** Distributional modeling for semantic shift detection. *International Journal of Lexicography*. 2019, vol. 32(2), pp. 163-183.

14**. Wool J.R.** Does the Cryptographic Hashing of Passwords Qualify for Statutory Breach Notification Safe Harbor. *Journal of Law & Cyber Warfare*. 2018, vol. 6(2), pp. 56-89.

15. **Zhu F., Fellbaum C.** Quantifying Fixedness and Compositionality in Chinese Idioms. *International Journal of Lexicography*. 2015,vol. 28(3), pp. 338-350.

16. **Beckwith R., Miller G.** Implementing a Lexical Network. *International Journal of Lexicography*, vol. 3(4), pp. 302-312.

17. **Tanyimboh T.T., Templeman A.B.** Calculating Maximum Entropy Flows in Networks. *The Journal of the Operational Research Society, New Research Directions*. 1993, vol. 44(4), pp. 383-396.

18. **Fontenelle T**. WordNet, FrameNet and Other Semantic Networks in the International Journal of Lexicography – The Net Result? *International Journal of Lexicography*. 2012, vol. 25(4), pp. 437-499.

19. **Gilquin G., Laporte S.** The Use of Online Writing Tools by Learners of English: Evidence From a Process Corpus. *International Journal of Lexicography*. 2021, vol. 34(4), pp. 472-492.

20. **Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J.** Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography*. 1990, vol. 3(4), pp. 235-244.

21. **Foxley E., Gwei G.M.** Synonymy and Contextual Disambiguation of Words, *International Journal of Lexicography*. 1989, vol. 2(2), pp. 111-134.

22. **Suyatinov S.I., Buldakova T.I., Vishnevskaya J.A.** Identification of Situations Based on Synergetic Model. *3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*. 2021.10.1109/SUMMA53307.2021.9632207.

23. **Dudnikov S., Miheev P., Grinkina T.** Evaluating of Word Embeddings Hyper-parameters of the Master Data in Russian-Language Information Systems. In: Hu, Z., Petoukhov, S., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics. CSDEIS 2019. Advances in Intelligent Systems and Computing 2020, vol. 1127. *Springer, Cham*. https://doi.org/10.1007/978-3-030-39216-1_7.

24. **Belyanova M., Chernobrovkin S., Latkin I., Gapanyuk Y**. Metagraph Based Approach for Neural Text Question Generation. In: van der Aalst, W.M.P., et al. Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. 2021, vol. 12602. *Springer, Cham*. https://doi.org/10.1007/978-3-030-72610-2_6.

25. **Kanev A., Terekhov V., Chernenky V., Proletarsky A**. Metagraph Knowledge Base and Natural Language Processing Pipeline for Event Extraction and Time Concept Analysis. *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. 2021. 10.1109/ElConRus51938.2021.9396541.

26. **Savchenko M., Tynchenko V**. Unsupervised Production Machinery Data Labeling Method Based on Natural Language Processing. *International Russian Smart Industry Conference (SmartIndustryCon)*. 2024. 10.1109/SmartIndustryCon61328.2024.10515763.

27**. Rogov A.A., Loukachevitch N.V.** Evaluating the Performance of Interpretability Methods in Text Categorization Task. *Lobachevskii J Math 45*. 2024, pp. 1234-1245.