Инженерный вестник Дона, №2 (2026)
ivdon.ru/ru/magazine/archive/n2y2026/10757

Метод извлечения семантических характеристик предложений на основе алгоритма нечеткой логики

Э.Р. Городецкий, Ю.М. Вишняков

Кубанский государственный университет, Краснодар

Аннотация: С быстрым ростом объема информации в Интернете, накоплением больших баз данных и постоянно поступающих сведений от различных датчиков и интеллектуальных систем, пользователям становится чрезвычайно сложно найти то, что они действительно ищут. Поэтому создание методов автоматического резюмирования считается очень важной задачей обработки естественного языка. Эти потребности стали стимулом к разработке различных методов и подходов извлечения смысловой и семантической информации из документов, ее классификации и систематизации. В статье разработана архитектура системы гибридно-синтаксического нечеткого извлечения семантических признаков из текста и представлена ее математическая формализация. Авторская методика позволяет перейти от эмпирических оценок важности слов к строгому формализованному исчислению их семантического веса.
Ключевые слова: семантика, предложение, извлечение, нечеткая логика, сравнение, данные.

За последние два десятилетия наблюдался стремительный прогресс в информационных науках и технологиях, что привело к сбору и накоплению огромного количества данных в самых разных областях. Ранее считалось, что хранение больших объемов данных может быть потенциально полезным в будущем. Однако в настоящее время признано, что «сырые» сведения сами по себе очень трудно понять человеку, что приводит к дилемме «богатство данных, бедность информации» [1]. В данном контексте в связи с экспоненциальным ростом количества и сложности источников информации становится все более важным предоставлять пользователям усовершенствованные механизмы для поиска точных данных в доступных документах. Резюмирование текста стало важным и своевременным инструментом, помогающим интерпретировать большие объемы текста. Процесс автоматического анализа и реферирования, при котором компьютер создает резюме длинного документа, значительно отличается от обработки текста человеком, поскольку человек может уловить и соотнести глубокий

смысл и темы текстовых документов. Автоматизация такого навыка очень сложна в реализации.

Таким образом, растущий объем данных привел к необходимости создания нового поколения вычислительных технологий и инструментов для извлечения семантических характеристик текста. Обработка естественного языка является краеугольным камнем на стыке лингвистики, информатики и искусственного интеллекта, революционизируя способ нашего взаимодействия с текстовыми данными и их анализа [2]. Эта междисциплинарная область не только стремится расшифровать лингвистические тонкости, но и ставит своей задачей наделить программно-аппаратные системы способностью понимать и генерировать язык, подобный человеческому.

Автоматическое резюмирование текста можно разделить на две категории в зависимости от подхода: i) резюмирование на основе абстракции и ii) резюмирование на основе извлечения. Большинство работ в этой области основано на последнем подходе, который использует нечеткие правила и нечеткие множества для выбора предложений на основе их характеристик. Методы нечеткой логики в форме приближенного рассуждения обеспечивают системы поддержки принятия решений мощными возможностями рассуждения [3]. Однако несмотря на то, что методы нечеткой логики довольно эффективны в извлечении семантических характеристик, сложность их использования связана с определением правил или базовых принципов, используемых при выводах. Кроме того, степени принадлежности обычного нечеткого множества, не способны обрабатывать различные типы неопределенностей, встречающиеся в высказываниях на естественном языке, например, субъективно выраженное мнение или описания числовых величин.

Таким образом, дальнейшее исследование данной проблематики имеет высокое научное и практическое значение, что и предопределило выбор темы

данной статьи.

Над разработкой метода лингвистического суммирования, основанного на алгоритмах нечеткой логики, который способен извлекать потенциально полезные и абстрактные знания как из числовых, так и из категориальных данных, трудятся Kanta Prasad Sharma, Mohd Shukri Ab Yajid [4], Zimanova D.A., Sadir A.K., Kassymov B.M. [5].

Особенности использования метода обобщения текста с использованием комбинации генетического алгоритма и генетического программирования для оптимизации наборов правил и функции принадлежности нечетких систем, описывают в своих публикациях Герасименко Е.М., Стеценко В.В. [6], Yanping Geng, Reem Alshahrani, Hana Mohammed Mujlid Enhancing [7].

Несмотря на активное развитие методов извлечения семантических характеристик текста на основе нечеткой логики, остаётся нерешённой проблема формализации и автоматической адаптации базы нечетких правил при переходе между предметными областями. Дополнительной сложностью является обеспечение устойчивости семантических оценок к языковой неоднозначности и шуму во входных данных при сохранении интерпретируемости выводов. Также недостаточно изучены вопросы масштабируемости таких алгоритмов при обработке больших корпусов текстов и их интеграции с современными нейросетевыми моделями без потери объяснимости.

Таким образом, цель статьи заключается в изучении подходов к извлечению семантических характеристик предложений на основе алгоритма нечеткой логики.

Согласно имеющимся на сегодняшний день наработкам, выделение семантических признаков на основе обработки естественного языка включает в себя несколько ключевых этапов. Начиная с заданного текстового корпуса, первым шагом является токенизация и предварительная обработка текста с

созданием матрицы термин-документ, где каждая строка соответствует документу, а каждый столбец представляет термин. Для семантического извлечения признаков может использоваться предварительно обученная модель вложения слов, например, такая как Word2Vec или GloVe, которая позволяет представлять термины в непрерывном векторном пространстве. Следующий этап включает создание матрицы сходства на основе векторов документов. Процесс иерархического кластеризации начинается с того, что каждый документ рассматривается как отдельный кластер. Посредством итеративной агломерации ближайшие кластеры объединяются на основе матрицы сходства. Матрица сходства обновляется после каждого слияния, обычно с использованием таких методов, как среднее связывание [8].

Анализ современных алгоритмов, используемых для извлечения семантических характеристик предложений (таких как «частотность терминов-обратная частотность документов», сообщение, описывающее локальное состояние маршрутизатора или сети, Линейный дискриминантный анализ, а также нейросетевых моделей трансформерного типа) позволяет отметить их следующие недостатки:

1. Игнорирование синтаксической структуры: статистические методы, основанные на частотности слов («мешок слов»), не учитывают грамматические связи внутри предложения. Для них подлежащее, несущее основную смысловую нагрузку, и второстепенное определение, встречающееся с той же частотой, математически эквивалентны, что приводит к потере контекста [9].

2. Проблема «жестких» порогов: классические алгоритмы используют бинарную логику отсечения (важно/не важно) на основе фиксированных пороговых значений. Однако естественный язык по своей природе нечеткий и градуальный: значимость слова — это не дискретная величина, а непрерывная степень уверенности.

3. Высокая ресурсоемкость и проблема интерпретируемости. Современные глубокие нейронные сети, несмотря на достаточно хорошие показатели точности классификации, характеризуются непрозрачностью процесса принятия решений (функционируют по принципу «черного ящика»). Внутренняя структура таких моделей не позволяет однозначно интерпретировать логику формирования результата, а процедура их обучения сопряжена со значительными вычислительными затратами [10].

4. Зависимость от словарей: методы, опирающиеся исключительно на онтологии, страдают от проблем покрытия: они плохо справляются с неологизмами, сленгом и узкоспециализированной терминологией.

Таким образом, для преодоления указанных сложностей и проблем автором разработана методика гибридно-синтаксического нечеткого извлечения семантических признаков (рис. 1). Ключевая идея заключается в интеграции разнородных факторов (статистики, синтаксической роли слова и его семантической полисемии) в единый вектор, обрабатываемый системой нечеткого вывода. В отличие от аналогов, методика позволяет взвешивать важность слов, имитируя логику эксперта-лингвиста.

Математическое описание системы гибридно-синтаксического нечеткого извлечения семантических признаков

1. Теоретико-графовая модель предложения (Блок I)

Пусть S — анализируемое предложение. Будем моделировать его как ориентированный ациклический граф синтаксических зависимостей:

$$G = (V, E, L)$$

где $V = \{w_1, w_2, \dots, w_m\}$ - множество вершин (слов/токенов);

$E \subseteq V \times V$ - множество ориентированных дуг (w_i, w_j), обозначающих зависимость слова w_j от w_i;

$L: E \to \Lambda$ — функция разметки дуг, где Λ = {nsubj, dobj, amod, ...} — множество типов синтаксических отношений.

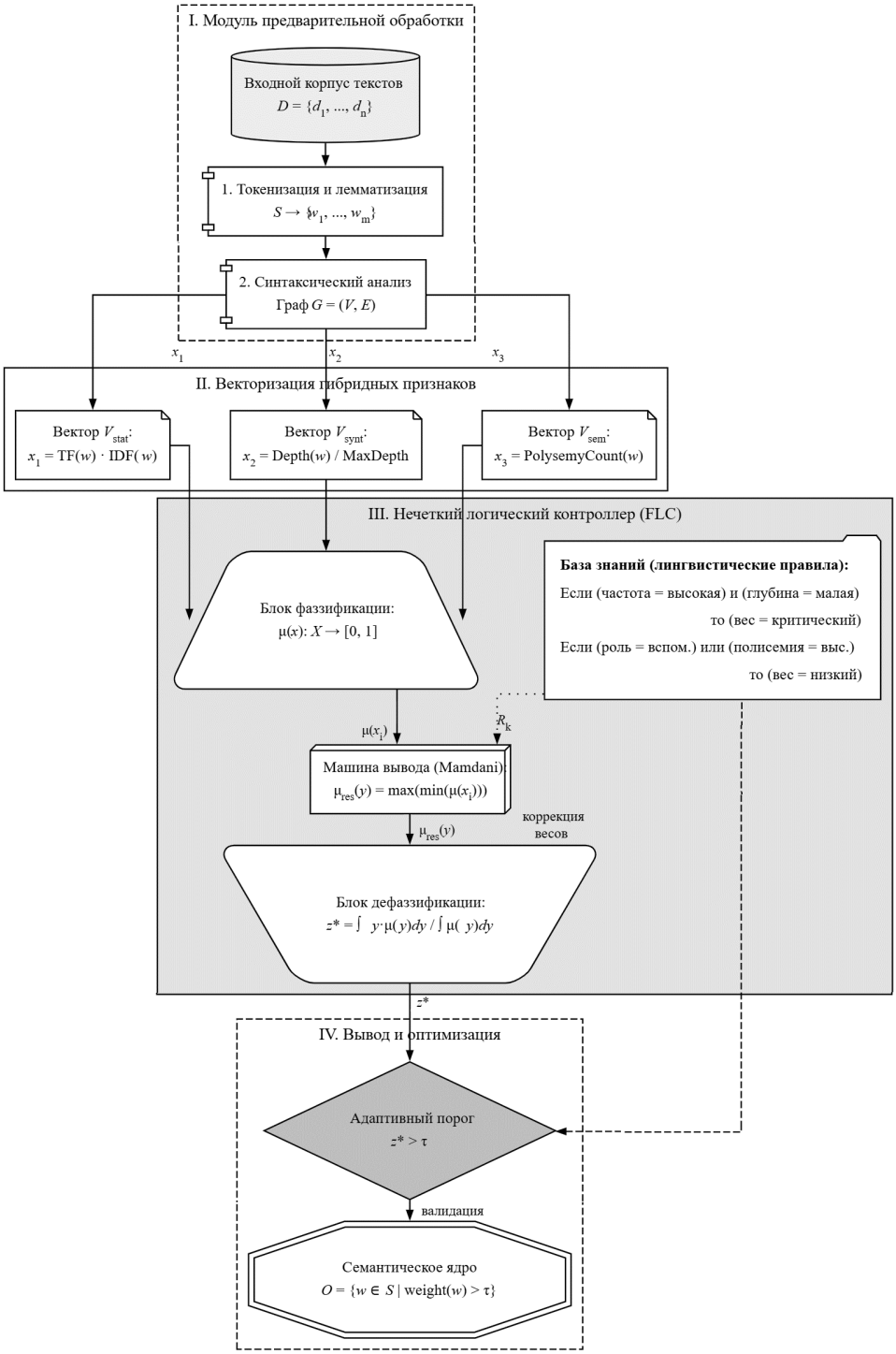

Рис. 1. Архитектура системы гибридно-синтаксического нечеткого извлечения семантических признаков (составлено автором)

Для каждого слова $w \in V$ определяется функция глубины $\delta(w)$,

представляющая собой длину кратчайшего пути от корня графа root до *w*:

$$\delta(w) = \begin{cases} 0 & \text{если } w = root \\ \delta\big(parent(w)\big) + 1 & \text{иначе} \end{cases}$$

2. Векторизация в гибридном пространстве признаков (Блок II)

Отображение слова в признаковое пространство $\Phi: V \to [0,1]^3$задается вектором $x = [x_1, x_2, x_3]^T$.

Статистический признак (x_1). Нормированная частота слова, взвешенная по обратной частоте в корпусе *D*:

$$x_1(w) = \frac{f_{w,S}}{max_{v \in S} f_{v,S}} \cdot \log\,(1 + \frac{|D|}{|\{d \in D: w \in d\}|}$$

где $f_{w,S}$ — частота слова *w* в предложении/тексте *S*.

Синтаксическая центральность (x_2). Чем выше слово в дереве разбора, тем оно важнее. Введем экспоненциальное затухание значимости с глубиной:

$$x_2(w) = \exp\,(-\lambda \cdot \frac{\delta(w)}{max_{v \in V}\delta(v)}$$

где $\lambda > 0$ — коэффициент затухания (настраиваемый гиперпараметр).

Семантическая определенность (x_3). Используя лексическую онтологию (WordNet) Ω определим *Syn(w)* как множество синсетов для леммы слова *w*. Мера однозначности:

$$x_3(w) = \frac{1}{\sqrt{1 + |Syn(w)|}}$$

3. Нечеткая система вывода (Блок III).

Для каждой входной переменной $x_j (j = 1 \ldots 3)$ определяется набор лингвистических термов $T_j = \{A_{j,1}, \ldots, A_{j,M_j}\}$. Используем дифференцируемые Гауссовы функции принадлежности (для возможности обучения):

$$\mu_{i,j}(x_j) = \exp\,(-\frac{(x_j - c_{i,j})^2}{2\sigma_{i,j}^2}$$

где $c_{i,j}$ — центр, а $\sigma_{i,}$— ширина i-го терма для j-го признака.

Агрегация правил. Пусть база знаний содержит K правил. Степень активации k-го правила рассчитывается через T-норму (произведение для сохранения гладкости функции):

$$w_k = \prod_{j=1}^{3} \mu_{k,j}(x_j)$$

Здесь $\mu_{k,j}$ — функция принадлежности, соответствующая терму во входной части k-го правила.

Логический вывод и импликация. Пусть y_k — четкое число, соответствующее заключению k-го правила (модель Такаги-Сугено 0-го порядка, упрощающая вычисления по сравнению с Мамдани, что предпочтительно для алгоритмизации):

$$R_k: \text{ЕСЛИ } (x_1 = \tilde{A}_k) \text{ и} (x_2 = \tilde{B}_k), \text{ТО } (y = \tilde{C}_k)$$

Дефаззификация. Итоговый семантический вес $z^*(w)$ вычисляется как взвешенное среднее:

$$z^*(w) = \frac{\sum_{k=1}^{K} w_k \cdot \theta_k}{\sum_{k=1}^{K} w_k}$$

4. Адаптивная оптимизация параметров (Блок IV).

Параметры системы не задаются жестко, а настраиваются. Пусть E — целевая функция ошибки на обучающей выборке (разница между предсказанным весом слова z^* и эталонной меткой эксперта t):

$$E = \frac{1}{2}(z^* - t)^2$$

Используем метод градиентного спуска для настройки параметров антецедентов (c, σ) и консеквентов (θ).

Правило обновления для центров функций принадлежности ($c_{i,j}$). Согласно правилу цепочки:

$$\frac{\partial E}{\partial c_{i,j}} = \frac{\partial E}{\partial z^*} \cdot \frac{\partial z^*}{\partial w_k} \cdot \frac{\partial w_k}{\partial \mu_{i,j}} \cdot \frac{\partial \mu_{i,j}}{\partial c_{i,j}}$$

Итоговая формула коррекции весов (η — скорость обучения):

$$c_{i,j}(t+1) = c_{i,j}(t) - \eta \cdot (z^* - t) \cdot \frac{\theta_k - z^*}{\sum w} \cdot w_k \cdot \frac{(x_j - c_{i,j})}{\sigma_{i,j}^2}$$

Таким образом, система самообучается: если слово было важным (синтаксически центральным), но система дала ему низкий вес, «колокол» функции принадлежности сдвигается в сторону значения признака этого слова.

5. Выходной фильтр. Финальное решение о включении слова в семантическое ядро принимается на основе α-среза:

$$Decision(w) = \mathbb{I}(z^*(w) \geq \tau_{adapt}$$

где $\tau_{adapt} = \mu_{glob} + \beta \cdot \sigma_{glob}$ - адаптивный порог, зависящий от среднего (μ_{glob}) и дисперсии (σ_{glob}) весов всех слов в документе.

Таким образом, подводя итоги проведенного исследования, можно сделать следующие выводы.

Цифровизация и быстрое развитие информационных технологий привело к созданию больших баз данных, хранящих разнообразные сведения в различных форматах и формах представления. Это существенным образом усложняет задачу анализа таких массивов информации, извлечения из них важных параметров и правильной интерпретации текста в целом. Учитывая актуальность отмеченной проблематики, в статье разработана авторская методика гибридно-синтаксического нечеткого извлечения семантических признаков, направленная на устранение недостатков традиционных частотных алгоритмов. В основе подхода лежит принцип многофакторной оценки лексем, объединяющий статистические показатели, синтаксическую роль слова и степень его полисемии в единый вектор признаков. Обработка сформированных векторов осуществляется посредством системы нечеткого

логического вывода, что позволяет моделировать когнитивные процессы эксперта-лингвиста.

Литература

1. Целых А.Н. Извлечение причинно-следственных кортежей из текста на основе глубокого обучения с использованием синтетических данных // Известия ЮФУ. Технические науки. 2025. № 1 (243). С. 118-129.

2. Герасименко Е.М., Стеценко В.В. Анализ тональности текстовых отзывов с применением тональных словарей и кардинальности нечеткого множества // Известия ЮФУ. Технические науки. 2022. № 5. С. 106-116.

3. Jain M., Jindal R., Jain A. Code-mixed Hindi-English text correction using fuzzy graph and word embedding // Expert Systems. 2023. Volume 41, Issue 7. pp. 98-105.

4. Sharma K. P., Shukri Ab Yajid M. A Systematic Review on Text Summarization: Techniques, Challenges, Opportunities // Expert Systems. 2025 Volume 42, Issue 4. pp. 34-41.

5. Zimanova D.A., Sadir A.K., Kassymov B.M. Использование нечетких множеств и логики для сентимент-анализа текста в социальных сетях // International Journal of Information and Communication Technologies. 2022. № 1(1). С. 34-41.

6. Герасименко Е.М., Стеценко В.В. Определение тональности текста с учетом влияния модификаторов интенсивности и применением нечетких правил // Известия ЮФУ. Технические науки. 2024. № 3 (239). С. 71-78.

7. Geng Ya., Alshahrani R., Mohammed Mujlid H. Enhancing Foreign Language Learning Through Social Media Applications: A Fuzzy Analytic Hierarchy Process Approach // European Journal of Education. 2025. Volume 60, Issue 1. pp. 87-94.

8. Вакушин А.А., Клебанов Б.И. Применение больших языковых

моделей в имитационном моделировании // Инженерный вестник Дона. 2024. №2. URL: ivdon.ru/ru/magazine/archive/n2y2024/8990 (дата обращения: 18.12.2025).

9. Syuhada Mohd Ali N., Suhairi Salleh I. Degumming and bleaching process troubleshooting in a palm oil refining process using fuzzy expert system with thematic analysis // Asia-Pacific Journal of Chemical Engineering. 2024. Volume 19, Issue 4. pp. 123-130.

10. Глухих И.Н., Глухих К.И. Разработка экспертных систем на основе большой языковой модели и генерации с дополненной выборкой // Инженерный вестник Дона. 2025. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2025/10496 (дата обращения: 18.12.2025).

References

1. Tselih A.N. Izvestiya YUFU. Tekhnicheskie nauki. 2025. No. 1 (243). pp. 118-129.

2. Gerasimenko E.M., Stetsenko V.V. Izvestiya YUFU. Tekhnicheskie nauki. 2022. No. 5. pp. 106-116.

3. Jain M., Jindal R., Jain A. Expert Systems. 2023. Volume 41, Issue 7. pp. 98-105.

4. Sharma K. P., Shukri Ab Yajid M. A Expert Systems. 2025. Volume 42, Issue 4. pp. 34-41.

5. Zimanova D. A., Sadir A. K., Kassymov B. M. International Journal of Information and Communication Technologies. 2022. No. 1 (1). pp. 34-41.

6. Gerasimenko E. M., Stetsenko V. V. Izvestiya YUFU. Tekhnicheskie nauki. 2024. No. 3 (239). pp. 71-78.

7. Geng Ya., Alshahrani R., Mohammed Mujlid H. European Journal of Education. 2025. Volume 60, Issue 1. pp. 87-94.

8. Vakushin A.A., Klebanov B.I. Inzhenernyj vestnik Dona. 2024. №2.

URL: ivdon.ru/ru/magazine/archive/n2y2024/8990 (date assessed 18.12.2025).

9. Syuhada Mohd Ali N., Suhairi Salleh I. Asia-Pacific Journal of Chemical Engineering. 2024. Volume 19, Issue 4. pp. 123-130.

10. Gluhih I.N., Gluhih K.I. Inzhenernyj vestnik Dona. 2025. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2025/10496 (data obrashcheniya: 18.12.2025).

Дата поступления: 11.12.2025

Дата публикации: 7.02.2026