Рубрика журнала
Научная статья
УДК 004.4'412

Особенности учета неопределенностей при машинной обработке естественных языков

М. В. Слюсарь, Р. Ссали*

Национальный исследовательский университет «МИЭТ», г. Москва, Россия

* *ssalironnie07@gmail.com*

Аннотация. При обработке естественного языка проблема неопределенностей – критическая задача для автоматизации на основе искусственного интеллекта в цифровой экономике. В работе раскрыты лексические, семантические, референциальные, прагматические неопределенности, возникающие в естественном языке. Предложен алгоритм, повышающий уровень надежности за счет разрешения контекстной неоднозначности и уменьшения количества неверных интерпретаций. Эмпирические результаты показали повышение точности классификации и лучшую адаптивность в неоднозначных сценариях. Предложенный подход может применяться в приложениях цифровой экономики, таких как чат-боты для обслуживания клиентов, анализ настроений для прогнозирования рынка и научных исследований. Показано, что при смягчении неопределенности алгоритм усиливает предиктивный контроль в автоматизированных системах на основе искусственного интеллекта.

Ключевые слова: управление неопределенностью, естественный язык, предиктивное управление, искусственный интеллект, лексико-смысловая неоднозначность, контекстная неоднозначность, цифровая экономика

Original article

Features of accounting for uncertainties in machine processing of natural languages

M. V. Sliusar, R. Ssali*

National Research University of Electronic Technology, Moscow, Russia

* *ssalironnie07@gmail.com*

Abstract. The problem of uncertainties in natural language processing is a critical challenge for AI-driven automation in the digital economy. In this work, lexical, semantic, referential, pragmatic uncertainties that occur in natural language processing are identified. An algorithm enhancing the reliability level due to context disambiguation and misinterpretations reduce is proposed. The empirical results have shown an improvement in classification accuracy and better adaptability in ambiguous scenarios. The proposed approach can be used in digital economy applications, such as customer service chat bots, sentiment analysis for market forecasting, and scientific research. It was demonstrated that by mitigating uncertainty, the algorithm strengthens AI-driven predictive control in automated systems.

Keywords: uncertainty management, natural language, predictive control, AI, lexical and semantic ambiguity, context ambiguity, digital economy

Введение. При обработке естественного языка возникают разные типы неопределенностей, которые ставят под сомнение надежность автоматизации на основе искусственного интеллекта (ИИ) в условиях цифровой экономики. Управление неопределенностями имеет решающее значение для улучшения

предиктивного управления в автоматизированных процессах, поскольку неправильная интерпретация может дестабилизировать производительность системы и повлиять на точность и доверие пользователей [1].

В настоящей работе рассматриваются существующие подходы к обработке неопределенностей и их ограничения, предлагается новый алгоритм для повышения точности и надежности обработки естественного языка (Natural Language Processing, NLP) с помощью ИИ за счет минимизации неоднозначностей языка.

Подходы к управлению неопределенностями в ИИ-ориентированном NLP. Типы неопределенностей в NLP:

– лексическая двусмысленность: слова с несколькими значениями, такие как «банк», зависят от контекста для точной интерпретации;

– семантическая двусмысленность: слова с одним значением, которое меняется в зависимости от контекста, например «свет» (освещенность и вес);

– референциальная неопределенность: трудность в определении ссылок на сущность, как, например, при выделении «он» во фразе «Джон и Мэри»;

– прагматическая двусмысленность: изменчивость предполагаемого значения в зависимости от тона, например «все в порядке» [2, 3].

Существуют разные подходы к решению проблемы неопределенностей NLP, каждый из которых способствует созданию надежных систем ИИ, хотя и с определенными ограничениями.

Статистические модели. Скрытые марковские модели (HMMs) и условные случайные поля (CRFs) используют вероятностные прогнозы для контекстной

деамбигуации. Однако они требуют больших вычислительных затрат и могут упускать особенности языка.

Системы, основанные на правилах. Эти системы опираются на предопределенные лингвистические правила для решения конкретных неоднозначных ситуаций, например:

– разрешение местоимения: такие правила, как «он», относятся к последнему субъекту мужского пола, уточняют обращения в “Valentin argued with Ronald. He left”. К кому относится «он»? Valentin or Ronald?

– грамматическая дезамбигуация: созданные вручную правила отмечают конфликтующие структуры. Например: “Time flies like an arrow”, т. е. время движется быстро (сравнение времени с летящей стрелой). Time = подлежащее, flies = глагол; “Fruit flies like a banana”, т. е. фруктовые мухи (насекомые) любят бананы. Fruit flies = подлежащее (словосочетание с существительным), like = глагол;

– чат-боты, ориентированные на конкретный домен: первые боты для обслуживания клиентов (например, в системах бронирования авиабилетов) используют правила, чтобы сопоставить фразу “I want to change my flight” с заранее определенным рабочим процессом.

Системы, основанные на правилах, эффективны для решения узких задач, они с трудом справляются с изменчивостью языка (например, сленгом, сарказмом) и плохо масштабируются на открытые контексты [1].

Подходы машинного обучения. Методы глубокого обучения, включая рекуррентные нейронные сети и трансформеры, позволяют улавливать сложные языковые паттерны и адаптироваться к контексту, но нередко требуют большого количества помеченных данных и вычислительных ресурсов [4, 5].

Вероятностные модели. Байесовские сети и марковские процессы принятия решений предлагают вероятностные рассуждения, но могут оказаться недостаточно масштабируемыми для решения масштабных задач NLP [6].

Разработка алгоритма для решения проблемы неопределенностей в NLP. Перечислим основные ограничения и проблемы существующих методов:

– зависимость от данных: методы машинного обучения опираются на большие наборы данных с метками, получение которых может быть дорогостоящим и отнимать много времени [7];

– контекстуальная сложность: многие методы с трудом справляются с интерпретацией тонких изменений контекста, что сказывается на точности при решении задач обнаружения сарказма [8];

– масштабируемость: высокие вычислительные требования, особенно для глубокого обучения, сужают применимость в условиях ограниченных ресурсов [9].

Предлагаемый алгоритм устраняет перечисленные ограничения, уравновешивая теоретическую строгость и практическую адаптивность, что повышает надежность приложений NLP.

В качестве примера разрешения лексических и семантических неоднозначностей рассмотрим английское предложение “The bank can guarantee deposits will be secure”. Слово “bank” неоднозначно и может относиться как к финансовому учреждению, так и к берегу реки. Для того чтобы разрешить эту лексическую неоднозначность, алгоритм использует следующие шаги [10]:

Шаг 1. Токенизация и извлечение контекста:

жетоны: $\{w_1,,,,,w_n\}$;

определитель: $\{d_1,,,,,d_n\}$;

контекст (c): {“The”, “can”, “guarantee”, “deposits”, “will”, “be”, “secure”} .

Шаг 2. Разграничение смыслов:

– возможный смысл: $(s)=\{f,r\}$;

– вычислить вероятность: $p(s/c)$. Алгоритм вычисляет вероятность каждого смысла с учетом наблюдаемого контекста, используя

$$p(s/c)=\frac{p(c/s)\cdot p(s)}{p(c)},$$

где $p(s/c)$ – вероятность того, что контексту будет придан определенный смысл; $p(s)$ – предварительная вероятность каждого смысла; $p(c)$ – вероятность наблюдаемого контекста.

Шаг 3. Выбор смысла: алгоритм выбирает смысл с наибольшей вероятностью $\text{sense}=\arg\text{Max}_s\, p(s/c)$.

Данный подход позволяет алгоритму использовать контекст для точного разрешения лексической неоднозначности слова “bank”, выбирая смысл с наибольшей контекстуальной релевантностью.

В качестве примера разрешения синтаксических двусмысленностей рассмотрим предложение: “The bank can guarantee deposits will be secure”. Алгоритм, разрешающий эту синтаксическую неоднозначность, будет учитывать контекст, построение дерева разбора, синтаксический анализ и правила предпочтения. Проанализируем два дерева разбора:

Дерево 1: (The (bank(can(guarantee(deposits will) be) secure))).

Дерево 2: ((The bank)(can(guarantee deposits)(will be secure))).

Затем применим правила предпочтения для выбора дерева разбора. Введем оценку предпочтений *ps* для каждого дерева разбора, основанную на синтаксических правилах и контексте:

$$ps(T) = R_w C_s .$$

Алгоритм выбирает дерево разбора с наивысшим баллом предпочтения:

$$p_t = \arg \mathrm{Max}_T \, ps(T) .$$

Таким образом, рассматривая интеграцию нейронных сетей и символьных рассуждений для обработки неопределенности в NLP,

$$F_p = (\alpha p_s) + (1 - \alpha) p_n ,$$

где F_p – комбинированное предсказание; α – весовой вектор, отражающий степень доверия к символическим рассуждениям $(0 \le \alpha \le 1)$; p_s – символическое предсказание; p_n – нейросетевое предсказание.

Полученные результаты эксперимента основаны на использовании моделей ByteNet + LSTM и ByteNet + ResNet, т. е. на сравнении предсказаний двух моделей (таблица).

Оценка точности экспериментальных моделей

Accuracy assessment of experimental models

Модель	Число итераций	Испытание, %	Обучение, %	Общая оценка, %
ByteNet + LSTM	25	98,3	96,8	95,2
ByteNet + ResNet	50	66,1	61,2	58,6

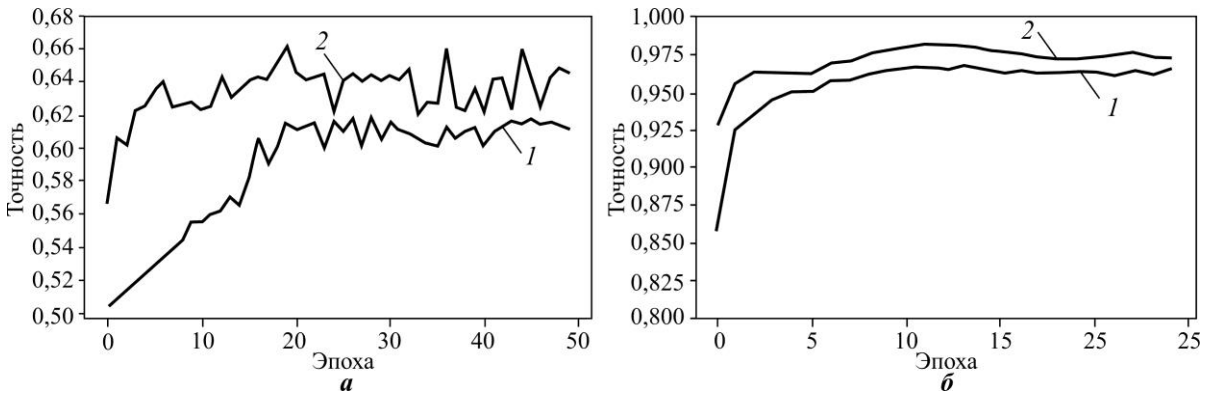
Точность рассчитывается по формуле

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)},$$

где TP – истинно положительные результаты; TN – истинно отрицательные результаты; FP – ложноположительные результаты; FN – ложноотрицательные результаты.

Приведенная формула измеряет долю правильно предсказанных экземпляров (как положительных, так и отрицательных) от общего числа экземпляров в наборе данных.

На рисунках показано, как меняется точность в процессе обучения и тестирования при использовании моделей ByteNet + ResNet и ByteNet + LSTM. Видно, насколько хорошо работает каждая модель и как улучшается или изменяется их точность с течением времени в процессе обучения и тестирования.



Визуализация точности моделей ByteNet + ResNet (*а*) и ByteNet + LSTM (*б*):

1 – точность обучения; *2* – тестовая точность

Visualization of the accuracy of the ByteNet + ResNet (*a*) and ByteNet + LSTM (*b*)

models: *1* – training accuracy; *2* – testing accuracy

Заключение. Предложенный алгоритм для решения проблемы неопределенностей в NLP улучшает контекстуальную деамбигуацию и вероятностные рассуждения. В результате эмпирической оценки модель ByteNet + LSTM на 30 % улучшила точность классификации по сравнению с базовой моделью ByteNet + ResNet с улучшенной производительностью при решении лексических, семантических и прагматических неоднозначностей. С качественной точки зрения алгоритм повышает интерпретируемость и устойчивость в практических задачах NLP, таких как анализ настроений и чат-боты, имеющих ключевое значение для приложений цифровой экономики – прогнозирования рыночных тенденций и автоматизации обслуживания клиентов. Это подтверждает способность гибридных архитектур противостоять неопределенностям, достигая при этом баланс между вычислительной эффективностью и масштабируемостью.

Литература / References

1. Поляков Д. В., Елисеев А. И., Андреев Д. С., Селиванов А. Ю. Подход к машинному обучению моделей открытых систем на основе теории нечетких множеств. In: *Цифровизация агропромышленного комплекса:* сб. статей II Междунар. науч.-практ. конф. (Тамбов, 21–23 окт. 2020). Т. 2. Тамбов: Тамбовский гос. техн. ун-т; 2020, с. 347–352. EDN: LSCBWD.

   Polyakov D. V., Eliseev A. I., Andreyev D. S., Selivanov A. Yu. A new approach to machine learning for fuzzy models of open systems. In: *Tsifrovizatsiya agropromyshlennogo kompleksa:* collected papers of 2nd International res.-to-pract. conf. (Tambov, 21–23 okt. 2020). Vol. 2. Tambov: Tambov State Tech. Univ.; 2020, pp. 347–352. (In Russ.).

2. Григорьев Д. В. Применение машинного обучения и естественного языкового процессинга в анализе текстовых данных: обзор методов, эффективности и перспектив. In: *Научные исследования молодых ученых:* сб. статей XXVIII Междунар. науч.-практ. конф. (Пенза, 27 июня 2024). Пенза: Наука и просвещение; 2024, с. 9–11. EDN: BDNYWX.

   Grigoriev D. V. Application of machine learning and natural language processing in text data analysis: Overview of methods, effectiveness and prospects. In: *Nauchnye issledovaniya molodykh uchenykh:* collected papers of 28th International res.-to-pract. conf. (Penza, 27 iyunya 2024). Penza: Nauka i prosveshchenie Publ.; 2024, pp. 9–11. (In Russ.).

3. Воробьева А. А., Федосенко М. Ю. Методы интеллектуального анализа данных и обработки естественного языка в управлении роботизированными производственными системами. *Доклады ТУСУР.* 2023;26(3):65–71. https://doi.org/10.21293/1818-0442-2023-26-3-65-71. EDN: UJGUUB.

   Vorobeva A. A., Fedosenko M. Yu. Methods for data mining and natural language processing in the management of robotic production systems. *Doklady TUSUR = Proceedings of the TUSUR University.* 2023;26(3):65–71. (In Russ.). https://doi.org/10.21293/1818-0442-2023-26-3-65-71

4. Бритов В. С., Мартышкин А. И., Данилов Е. А. Обзор и сопоставление в задачах обработки естественного языка сверхточных и рекуррентных нейронных сетей. *Тенденции развития науки и образования.* 2023;(97-12):41–45. https://doi.org/10.18411/trnio-05-2023-654. EDN: JBPZQT.

   Britov V. S., Martyshkin A. I., Danilov E. A. Review and comparison in natural language processing tasks of convolutional and recurrent neural networks. *Tendentsii razvitiya*

nauki i obrazovaniya. 2023;(97-12):41–45. (In Russ.). https://doi.org/10.18411/trnio-05-2023-654

5. Лобач В. И., Зайцева А. В., Лобач С. В. О байесовском подходе в машинном обучении на основе MCMC-метода. *Экономика, моделирование, прогнозирование.* 2021;(15):192–198. EDN: IGWWRP.

   Lobach V., Zaytseva A., Lobach S. On bayesian approach in machine learning based on MCMC-method. *Ekonomika, modelirovanie, prognozirovanie.* 2021;(15):192–198. (In Russ.).

6. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D. et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Boston, MA: IEEE; 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

7. Гудков А. А. Прогнозирование эффективности проектной деятельности на основе интеграции подходов бизнес-аналитики и машинного обучения. *Вестник Волжского университета им. В. Н. Татищева.* 2024;2(1):27–36. https://doi.org/10.51965/2076-7919_2024_2_1_27. EDN: QYXZHC.

   Gudkov A. A. Forecasting the efficiency of project activities based on the integration of business analytics and machine learning approaches. *Vestnik Volzhskogo universiteta im. V. N. Tatishcheva = Reporter of Volzhsky University after V. N. Tatischev.* 2024;2(1):27–36. (In Russ.). https://doi.org/10.51965/2076-7919_2024_2_1_27

8. Brownlee J. Encoder-decoder recurrent neural network models for neural machine translation. *Machine Learning Mastery.* 07.08.2019. Available at: https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/ (accessed: 10.09.2025).

9. Алейник А. О. Применение нейронных сетей для обработки естественного языка в информационных технологиях. *Студенческий.* 2023;(26-1):17–18. EDN: TKEYHQ.

Aleynik A. O. Neural networks application for natural language processing in information technologies. *Studencheskiy.* 2023;(26-1):17–18. (In Russ.).

10. Капелько И. В., Шиман Д. В. Трехуровневая токенизация для автоматического реферирования текста. In: *Информационные технологии:* материалы 83-й науч.-техн. конф. ППС, науч. сотрудников и аспирантов (с междунар. участием) (Минск, 04–15 февр. 2019). Минск: БГТУ; 2019, с. 53. EDN: YZVFNZ.

Kapel’ko I. V., Shiman D. V. Three-level tokenization for automated text summarization. In: *Informatsionnye tekhnologii:* proceedings of 83th sci.-tech. conf. for academic teaching staff, scientific researchers and postgraduate students (with international participation) (Minsk, 04–15 fevr. 2019). Minsk: BelSTU; 2019, p. 53. (In Russ.).

Информация об авторах

Слюсарь Мария Владимировна – аспирант, ассистент Института системной и программной инженерии и информационных технологий Национального исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, 1), mar_vin@bk.ru

Ссали Рональд – аспирант Института системной и программной инженерии и информационных технологий Национального исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, 1), ssalironnie07@gmail.com

Information about the authors

Maria V. Sliusar – PhD student, Assistant of the Institute of System and Software Engineering and Information Technologies, National Research University of Electronic Technology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), mar_vin@bk.ru

Ronald Ssali – PhD student of the Institute of System and Software Engineering and Information Technologies, National Research University of Electronic Technology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), ssalironnie07@gmail.com