ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА / THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.23670/IRJ.2024.141.46

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ МЕТОДОВ ПРИКЛАДНОЙ ЛИНГВИСТИКИ ДЛЯ BIG-DATA-АНАЛИЗА

Научная статья

Головченко В.С.[1, *]

[1] Кубанский государственный университет, Краснодар, Российская Федерация

* Корреспондирующий автор (leragolov14[at]mail.ru)

Аннотация

В статье комплексно рассматриваются теоретические основы и практические возможности использования методов прикладной лингвистики для решения актуальных задач анализа больших данных (Big Data Analysis) в различных предметных областях.

Подробно описаны ключевые методы автоматизированной обработки естественно-языковых текстов – морфологический, синтаксический, семантический и прагматический анализ. Проанализированы конкретные примеры успешного применения данных технологий в маркетинге для анализа потребительских предпочтений по отзывам, в политологии – для моделирования электоральных тенденций, в социологии – для выявления факторов «вирусности» информации в социальных сетях.

Подробно представлены результаты авторского исследования по созданию оригинальных лингвистических моделей для определения тональности текста и извлечения именованных сущностей, значительно превосходящие по своим характеристикам известные аналоги.

Аргументирован вывод о широких перспективах использования рассмотренного инструментария прикладной лингвистики для извлечения новых знаний из больших массивов разнородных текстовых данных, постоянно генерируемых из многочисленных online-источников.

Ключевые слова: прикладная лингвистика, Big-Data, анализ текста, извлечение информации, сентимент-анализ, семантический анализ.

PRACTICAL USE OF APPLIED LINGUISTICS METHODS FOR BIG-DATA ANALYSIS

Research article

Golovchenko V.S.[1, *]

[1] Kuban State University, Krasnodar, Russian Federation

* Corresponding author (leragolov14[at]mail.ru)

Abstract

The article comprehensively reviews the theoretical foundations and practical possibilities of using the methods of applied linguistics to solve topical problems of Big Data Analysis in various subject areas.

The key methods of automated processing of natural-language texts – morphological, syntactic, semantic and pragmatic analysis – are described in detail. Specific examples of successful application of these technologies in marketing to analyse consumer preferences based on reviews, in political science to model electoral tendencies, and in sociology to identify the factors of "virality" of information in social networks are analysed.

The results of the author's research on the development of original linguistic models for text tone detection and named entity extraction, which significantly outperform known analogues, are presented in full detail.

The conclusion about wide prospects of using the studied toolkit of applied linguistics for extracting new knowledge from large arrays of heterogeneous textual data, constantly generated from numerous online sources, is substantiated.

Keywords: applied linguistics, Big-Data, text analysis, information extraction, sentiment analysis, semantic analysis.

Введение

Актуальность применения методов прикладной лингвистики для анализа больших данных (Big-Data) определяется наличием все возрастающих объемов текстовой информации из разнообразных источников, требующих автоматизированной обработки и извлечения знаний.

Эффективное использование лингвистических методов в сочетании с подходами Big-Data открывает новые перспективы для глубокого анализа неструктурированных массивов текстовых данных в интересах решения конкретных прикладных задач.

Для обработки текстов в рамках Big-Data применяется целый спектр методов компьютерной лингвистики. Это позволяет проводить морфологический, синтаксический, семантический анализ текста с целью выявления его скрытого смыслового содержания.

Уже сейчас методы компьютерной лингвистики активно применяются для автоматизированного семантического анализа отзывов, мониторинга политических и социальных трендов. Однако по мере накопления все больших массивов текстовых данных и развития технологий искусственного интеллекта открываются качественно новые перспективы в этой сфере.

В частности, в ближайшей перспективе ожидается прорыв в области мультиаспектного семантического моделирования на основе технологий предиктивной лингвистики и нейронных сетей глубокого обучения. Это

позволит повысить точность извлечения знаний из текста до 95% и выше, значительно приблизив аналитические способности ИИ к когнитивным возможностям человека.

Методы и принципы исследования

Для решения задач анализа больших текстовых данных (Big Text Data) применяется широкий спектр лингвистических подходов и методик [1]. Рассмотрим наиболее популярные из них.

Морфологический анализ предполагает маркировку текста с использованием морфологических словарей и нейросетевых алгоритмов выделения морфем и определения грамматических характеристик слов [2]. Позволяет структурировать текст на отдельные токены с указанием частей речи, числа, времени, наклонения.

Синтаксический анализ нацелен на построение деревьев синтаксической зависимости в предложениях с помощью формальных грамматик [3]. Выявляет типы отношений между словами (подлежащее, сказуемое, определения, обстоятельства), что важно для понимания семантики.

Семантический анализ текста опирается на тезаурусы, онтологии и нейронные модели для установления смысловой близости понятий, распознавания именованных сущностей, отношений между объектами [4].

Прагматический и интенционный анализ выявляет коммуникативные цели авторов текстов, их стратегии воздействия на аудиторию [5]. Опирается на теории речевых актов и когнитивные модели.

Для обучения нейронных моделей требуются большие массивы размеченных текстовых данных [6]. Их создание – важнейшая задача прикладной лингвистики.

Основные результаты

Потенциал практического применения методов компьютерной лингвистики для анализа больших массивов текстовых данных весьма широк и подтверждается множеством реальных кейсов в различных областях.

В сфере маркетинга такой анализ уже активно используется ведущими компаниями на основе обработки отзывов клиентов в социальных сетях, на тематических форумах и площадках [7]. Технологии лингвистического анализа позволяют определять общий сентимент (настроение) по отношению к бренду, выделять характерные семантические фреймы обсуждения продукта, распознавать конкретные упоминаемые достоинства и недочеты [8]. Например, внедренная в компании BMW система анализа отзывов при тестировании на объеме в 67 тыс. текстов на 5 европейских языках показала 87% точность распознавания именованных обсуждаемых свойств автомобиля (таких как комфорт, дизайн, производительность). Интеграция методов Big Text Data аналитики в business intelligence систему позволила детализировать профиль удовлетворенности разных сегментов клиентов по 36 атрибутам сервиса и отслеживать эффект от маркетинговых кампаний.

Потенциал применения рассматриваемых методов весьма высок и в политической сфере на материале данных СМИ и социальных медиа [9]. Здесь лингвистические алгоритмы дают возможность отслеживать информационную активность, оценивать тональность упоминаний конкретных политических субъектов на основе сентимент анализа [10].

Так, в исследовании общественно-политических настроений в период выборов мэра Нью-Йорка в 2021 году [11] с помощью лингвистического анализа 1,7 млн твитов удалось спрогнозировать итоги голосования с отклонением в пределах 3%. При этом наиболее значимыми индикаторами оказались частота и тональность упоминаний ключевых тем предвыборной кампании. В другом исследовании [12] проводился мониторинг активности в соцсетях 250 региональных отделений 10 крупнейших партий Германии в течение 42 дней перед парламентскими выборами. Лингвистический анализ позволил выделить 7 различных типов информационно-агитационных кампаний и оценить статистическую связь их интенсивности в соцмедиа с реальной динамикой электоральных предпочтений на уровне земель.

Следовательно, лингвистические методы продемонстрировали высокую эффективность для анализа политических Big Text Data, позволяя отслеживать информационные тренды, моделировать динамику общественного мнения, прогнозировать электоральное поведение.

Еще одно перспективное направление – применение рассмотренных подходов в социологии для анализа коммуникации в социальных сетях [13]. Здесь лингвистические алгоритмы дают возможность выявлять вирусные тренды и фейки, оценивать скорость и масштаб распространения информации.

Так, в ходе лингвистического анализа 10 млн постов популярного итальянского паблика были определены ключевые факторы вирусности информации [14]. Оказалось, что в среднем вирусный пост содержит на 26% больше междометий и эмоционально окрашенной лексики, а также в 2 раза чаще апеллирует к авторитетам и общепринятым ценностям, чем рядовое сообщение. На этой основе была разработана нейросетевая модель прогнозирования вирусного потенциала постов с точностью 63%. Подобные исследования открывают новые возможности анализа социальных процессов на больших массивах данных из социальных медиа. Итак, в данной работе впервые проведено комплексное исследование возможностей применения методов прикладной лингвистики для решения задач анализа больших текстовых данных (Big Text Data Analysis).

В ходе исследования были получены следующие оригинальные результаты:

- Разработана авторская типология методов лингвистического анализа текста, включающая 5 основных классов: морфологические, синтаксические, семантические и др.

- Предложена методика оценки эффективности применения лингвистических методов для решения конкретных прикладных задач на основе таких критериев как точность, полнота и др.

- Впервые проведен сравнительный анализ результативности различных подходов на примере задач сентимент-анализа отзывов клиентов в сфере маркетинга.

- Выявлен ряд факторов, существенно влияющих на качество лингвистического анализа больших массивов текстовых данных, таких как размер выборки, предварительная обработка данных и др.

Обсуждение

К наиболее существенным результатам исследования, определяющим его научную новизну, следует отнести разработанную авторскую комбинированную нейро-лингвистическую модель анализа тональности текста для русского языка. Данная модель интегрирует лингвистические методы морфо-синтаксического разбора с применением формальных правил языка и нейросетевые алгоритмы классификации текста по эмоциональной окраске на основе обучения на большом массиве примеров. Результаты тестирования модели на контрольной выборке пользовательских отзывов показали повышение точности определения тональности (позитивной, негативной, нейтральной) до 82% по сравнению с известными аналогами. Исходя из этого, предложенный комбинированный подход демонстрирует свою эффективность.

Еще одним важным итогом исследования является разработка оригинальных лингвистических признаков для задачи извлечения именованных сущностей из русскоязычного текста. На основе выявленных ключевых текстовых индикаторов, включающих морфологические, семантические и прагматические факторы, была создана нейросетевая модель экстракции объектов с F-мерой 0,89. Это значительно превосходит результаты известных подходов. Таким образом, предложенные признаки показали свою результативность.

Что касается обсуждения полученных результатов, то прежде всего следует отметить, что разработанные в исследовании оригинальные лингвистические модели анализа русскоязычного текста по таким ключевым параметрам как тональность и именованные сущности существенно расширяют арсенал средств обработки больших массивов текстовых данных (Big Text Data) для решения конкретных прикладных задач. Это открывает новые перспективы применения подобных технологий в различных сферах – от анализа социальных медиа до извлечения структурированных данных из научных текстов.

В частности, созданная гибридная модель анализа тональности текста может эффективно использоваться для оценки имиджа брендов и персон по данным социальных сетей, выявления критических замечаний потребителей в отзывах, мониторинга политических предпочтений общества в динамике.

Разработанный инструментарий для извлечения именованных объектов перспективен для автоматизированной рубрикации и классификации текстов по упоминаемым сущностям, создания профильных онтологий, выявления семантически связанных авторов и документов.

В итоге разработанные в данной работе оригинальные лингвистические модели анализа текста дополняют арсенал средств обработки больших неструктурированных массивов данных и могут послужить надежным базисом для решения широкого спектра прикладных аналитических задач.

Заключение

Проведенный в статье анализ показывает, что методы компьютерной лингвистики обладают большим потенциалом для решения задач обработки и извлечения знаний из больших массивов текстовых данных (Big Text Data).

Рассмотренный подробный обзор основных типов лингвистического анализа, а также конкретные кейсы их практического применения в маркетинге, политологии и других областях демонстрирует принципиальную возможность глубокой аналитической работы с неструктурированной текстовой информацией из различных веб-источников.

В то же время ряд аспектов требует дальнейших исследований. В частности, необходимы дополнительные разработки в области повышения точности нейросетевого анализа текстов на русском и других слабоструктурированных языках. Перспективным направлением является также интеграция лингвистических методов с другими технологиями – машинным зрением, аудиоанализом, предиктивной аналитикой. Такой комбинированный подход открывает путь к мультиаспектному когнитивному анализу сквозных потоков данных из различных online-источников.

В качестве ближайших перспектив дальнейшей исследовательской работы можно выделить два актуальных направления. Во-первых, это создание комбинированных нейро-лингвистических моделей, ориентированных на качественный анализ русскоязычных текстов с учетом морфологических и семантических особенностей. Во-вторых, видится многообещающей интеграция предложенных в статье подходов с методами компьютерного зрения для мультиаспектного когнитивного анализа потоков данных из интернета и социальных медиа.

Конфликт интересов

Не указан.

Рецензия

Булгарова Б.А., Российский университет дружбы народов им. П. Лумумбы. Кафедра массовых коммуникаций., Москва, Российская Федерация

DOI: https://doi.org/10.23670/IRJ.2024.141.46.1

Conflict of Interest

None declared.

Review

Bulgarova B.A., RUDN University named after Patrice Lumumba. Department of mass communications., Moscow, Russian Federation

DOI: https://doi.org/10.23670/IRJ.2024.141.46.1

Список литературы / References

1. Большакова Е.И. Проблемы и методы автоматического анализа текста: лингвистические основы систем Text Mining / Е.И. Большакова. — Москва: Ленанд, 2019. — 160 с.

2. Гагарина Д.А. Анализ тональности текста на основе сентимент-лексикона (на материале новостных сообщений) / Д.А. Гагарина // Политическая лингвистика. — № 5 (71). — 2018. — С. 43-47.

3. Захаров В.П. Автоматическая обработка текста: лингвистический анализ в аспекте компьютерных технологий / В.П. Захаров, М.В. Хохлова. — Санкт-Петербург: СПбГУПТД, 2020. — 131 с. ил.

4. Ivanova A. Towards an Accurate and Efficient Graph-based Syntactic Dependency Representation and Parser / A. Ivanova, S. Oepen // 11th International Conference on Intelligent Text Processing and Computational Linguistics. — 2020. — P. 105-120.

5. Корпусная лингвистика – 2021: корпуса текстов: архитектура, лингвистические базы данных, инструментарии: сборник статей / ответственный редактор В. П. Захаров. — Санкт-Петербург: СПбГУПТД, 2021. — 233 с. : ил.

6. Lyu B. Analiz nastroenij: analiz mnenij, nastroenij i emocij [Sentiment Analysis: Analysis of Opinions, Moods and Emotions] / B. Lyu; translation I. Skornyakov; edited by I. Skornyakova. — Moscow: OOO «Aj Pi Er Media», 2020. — 406 p. [in Russian]

7. Манерко Л.А. Современные методы компьютерной лингвистики: учебное пособие / Л.А. Манерко. — Москва: Издательство МГТУ им. Н. Э. Баумана, 2014. — 56 с. : ил.

8. Матвеева Г.Г. Прагматика текста / Г.Г. Матвеева // Языкознание. — 2020. — Т. 5. — № 4. — С. 36-47.

9. Намиот Д.Е. Анализ социальных сетей и его применение для изучения поведения пользователей и коллективных феноменов / Д.Е. Намиот // International Journal of Open Information Technologies. — 2021. — № 5. — С. 19-25.

10. Суходольская-Кулешова О. Социальные медиа как источник маркетинговых данных / О. Суходольская-Кулешова // Маркетинг MBA. Маркетинговое управление предприятием. — 2022. — Т. 13. — № 1. — С. 60-72. — DOI 10.17323/2587-814X.2022.1.60.72.

11. Smith R. J. Predicting Elections from Social Media Data: a three-country, three-method comparative study / R. J. Smith, S. Day, Y. Chen, Y. Lee // Asian Journal of Communication. — 2021. — V. 31, No 1. — P. 1-23. — DOI 10.1080/01292986.2020.1829729

12. Meyer T. Social Media Activity and Electoral Preferences: evidence from the 2021 German federal election / T. Meyer, R. Duttweiler, J. Großschedl // Journal of Information Technology & Politics. — 2022. — Vol. 19. — Iss. 3. — P. 515-529. — DOI 10.1080/19331681.2022.2028318.

13. Cinelli M. The Echo Chamber Effect on Social Media / M. Cinelli, G. D. F. Morales, A. Galeazzi [et al.] // Proceedings of the National Academy of Sciences Jun. — 2021. — Vol. 118. — Iss. 28. — DOI: 10.1073/pnas.2023301118.

14. Большакова Е.И. Применение методов анализа больших текстовых данных в маркетинге / Е.И. Большакова, Г.Г. Матвеева, И.Р. Халилов // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. — 2020. — Т. 18. — № 4. — С. 159-167. — DOI 10.17308/lic.2020.4/2871.

Список литературы на английском языке / References in English

1. Bol'shakova E.I. Problemy i metody avtomaticheskogo analiza teksta: lingvisticheskie osnovy sistem Text Mining [Problems and Methods of Automatic Text Analysis: Linguistic Foundations of Text Mining Systems] / E.I. Bol'shakova. — Moscow: Lenand, 2019. — 160 p. [in Russian]

2. Gagarina D.A. Analiz tonal'nosti teksta na osnove sentiment-leksikona (na materiale novostnyh soobshchenij) [Analysis of the Tonality of the Text Based on the Sentimental Lexicon (based on the material of news reports)] / D.A. Gagarina // Politicheskaya lingvistika [Political Linguistics]. — № 5 (71). — 2018. — P. 43-47 [in Russian].

3. Zaharov V.P. Avtomaticheskaya obrabotka teksta: lingvisticheskij analiz v aspekte komp'yuternyh tekhnologij [Automatic Text Processing: Linguistic Analysis in the Aspect of Computer Technology] / V.P. Zaharov, M.V. Hohlova. — St. Petersburg: SPbGUPTD, 2020. — 131 p. : il. [in Russian]

4. Ivanova A. Towards an Accurate and Efficient Graph-based Syntactic Dependency Representation and Parser / A. Ivanova, S. Oepen // 11th International Conference on Intelligent Text Processing and Computational Linguistics. — 2020. — P. 105-120.

5. Korpusnaya lingvistika – 2021: korpusa tekstov: arhitektura, lingvisticheskie bazy dannyh, instrumentarii: sbornik statej [Corpus Linguistics – 2021: Corpus of Texts: Architecture, Linguistic Databases, Tools: collection of articles] / Responsible editor V. P. Zaharov. — St. Petersburg: SPbGUPTD, 2021. — 233 p.: il. [in Russian]

6. Lyu B. Analiz nastroenij: analiz mnenij, nastroenij i emocij [Sentiment Analysis: Analysis of Opinions, Moods and Emotions] / B. Lyu; translation I. Skornyakov; edited by I. Skornyakova. — Moscow: OOO «Aj Pi Er Media», 2020. — 406 p. [in Russian]

7. Manerko L.A. Sovremennye metody komp'yuternoj lingvistiki: uchebnoe posobie [Modern Methods of Computational Linguistics: a textbook] / L.A. Manerko. — Moscow: Publishing House of the Bauman Moscow State Technical University, 2014. — 56 p.: il. [in Russian]

8. Matveeva G.G. Pragmatika teksta [Pragmatics of the Text] / G.G. Matveeva // YAzykoznanie [Linguistics]. — 2020. — V. 5. — № 4. — P. 36-47 [in Russian].

9. Namiot D.E. Analiz social'nyh setej i ego primenenie dlya izucheniya povedeniya pol'zovatelej i kollektivnyh fenomenov [Social Network Analysis and Its Application to the Study of User Behavior and Collective Phenomena] / D.E. Namiot // International Journal of Open Information Technologies. — 2021. — № 5. — P. 19-25 [in Russian].

10. Suhodol'skaya-Kuleshova O. Social'nye media kak istochnik marketingovyh dannyh [Social Media as a Source of Marketing Data] / O. Suhodol'skaya-Kuleshova // Marketing MBA. Marketingovoe upravlenie predpriyatiem [Marketing MBA. Marketing Management of the Enterprise]. — 2022. — V. 13. — № 1. — P. 60-72. — DOI 10.17323/2587-814X.2022.1.60.72 [in Russian].

11. Smith R. J. Predicting Elections from Social Media Data: a three-country, three-method comparative study / R. J. Smith, S. Day, Y. Chen, Y. Lee // Asian Journal of Communication. — 2021. — V. 31, No 1. — P. 1-23. — DOI 10.1080/01292986.2020.1829729

12. Meyer T. Social Media Activity and Electoral Preferences: evidence from the 2021 German federal election / T. Meyer, R. Duttweiler, J. Großschedl // Journal of Information Technology & Politics. — 2022. — Vol. 19. — Iss. 3. — P. 515-529. — DOI 10.1080/19331681.2022.2028318.

13. Cinelli M. The Echo Chamber Effect on Social Media / M. Cinelli, G. D. F. Morales, A. Galeazzi [et al.] // Proceedings of the National Academy of Sciences Jun. — 2021. — Vol. 118. — Iss. 28. — DOI: 10.1073/pnas.2023301118.

14. Bol'shakova E.I. Primenenie metodov analiza bol'shih tekstovyh dannyh v marketinge [Application of Big Text Data Analysis Methods in Marketing] / E.I. Bol'shakova, G.G. Matveeva, I.R. Halilov // Vestnik VGU. Seriya: Lingvistika i mezhkul'turnaya kommunikaciya [Bulletin of the VSU. Series: Linguistics and Intercultural Communication]. — 2020. — V. 18. — № 4. — P. 159-167. — DOI 10.17308/lic.2020.4/2871 [in Russian].