УДК 004.85

Обработка текста методами естественного языка

Л. В. Рудак, С. А. Зори
E-mail: semerikov2917@yandex.ru, ik.ivt.rec@mail.ru

Аннотация
В работе рассматриваются методы обработки текста с использованием естественного языка (NLP), которые играют ключевую роль в современном мире информационных технологий. Статья охватывает основные концепции и техники NLP, такие как токенизация, стемминг, лемматизация, удаление стоп-слов, использование регулярных выражений, а также методы представления текста, включая Bag of Words и TF-IDF. Особое внимание уделено анализу тональности, машинному переводу, автоматическому резюмированию и чат-ботам, которые являются важными направлениями в области NLP.

Введение

Обработка естественного языка является одной из ключевых областей искусственного интеллекта, направленной на взаимодействие между компьютерами и людьми с использованием естественного языка. В последние годы, благодаря достижениям в области машинного обучения и росту вычислительных мощностей, методы NLP значительно продвинулись вперед, предоставляя новые возможности для анализа, интерпретации и генерации текстовых данных [1].

Основная цель NLP заключается в создании систем, которые могут понимать и генерировать человеческий язык таким образом, чтобы это было полезно для различных прикладных задач, таких как автоматическое резюмирование текстов, машинный перевод, распознавание речи, анализ тональности, классификация текстов и многое другое. Эти задачи требуют использования различных методов и алгоритмов, которые могут обрабатывать и анализировать огромные объемы текстовых данных.

Методы NLP опираются на широкий спектр технологий и подходов, включая токенизацию, стемминг, лемматизацию, удаление стоп-слов, векторизацию текста и использование сложных моделей глубокого обучения. Одним из ключевых аспектов обработки текста является преобразование неструктурированных текстовых данных в структурированную форму, которая может быть использована для дальнейшего анализа и обработки. Это преобразование включает в себя извлечение значимой информации, устранение избыточности и учет контекста слов в тексте [2].

Таким образом, развитие методов обработки текста с помощью NLP продолжает активно трансформировать способы взаимодействия человека с информацией, делая этот процесс более эффективным и интуитивным.

Основные концепции и методы NLP

Процесс обработки языка с помощью NLP включает несколько ключевых этапов, таких как предобработка текста, синтаксический и семантический анализ, извлечение сущностей, анализ намерений и генерация ответов. Каждый из этих этапов играет важную роль в обеспечении точности и эффективности систем обработки информации. В предобработке текста важными шагами являются токенизация, нормализация и удаление стоп-слов, что позволяет структурировать и очистить исходные данные. Синтаксический анализ направлен на понимание грамматической структуры текста, а семантический анализ — на интерпретацию его смысла.

Извлечение сущностей (Named Entity Recognition, NER) и анализ намерений (Intent Recognition) являются одними из наиболее важных задач в NLP, так как они помогают системе понимать, какие объекты и действия упоминаются в тексте. Наконец, генерация ответов (Text Generation) позволяет системе формулировать ответы на запросы пользователей на естественном языке, что делает взаимодействие с системой более дружественным и эффективным.

Токенизация

Токенизация предложений — это процесс разделения строки текста на составляющие предложения. Задача на первый взгляд является не самой сложной, потому что в русском и многих других языках можно разделять предложения всякий раз, когда видим знак препинания. Токенизация слов — это процесс разделения строки текста на составляющие ее слова.

Токенизация является фундаментальным шагом в процессе обработки естественного языка. Выбор метода токенизации зависит от конкретной задачи и характеристик текста. Программные методы, такие как регулярные выражения, и специализированные библиотеки на языке Python, такие как NLTK и SpaCy, предоставляют широкий спектр инструментов для эффективной токенизации слов и предложений, что значительно упрощает дальнейший анализ и обработку текстовых данных [3].

Стемминг и лемматизация

В обработке естественного языка стемминг и лемматизация являются важными методами для нормализации текста. Они помогают привести слова к их базовым или корневым формам, что упрощает дальнейший анализ и обработку данных [4].

Стемминг — это процесс уменьшения слова до его основы или корня (стема), удаляя суффиксы и окончания. Это простая и быстрая техника, которая часто применяется в задачах извлечения информации и поиска данных.

Лемматизация — это процесс приведения слова к его канонической форме (лемме), используя знания о морфологии и части речи слова. В отличие от стемминга, лемматизация более сложная и точная, так как учитывает контекст и грамматические характеристики слова.

Стемминг и лемматизация являются важными инструментами в арсенале методов предобработки текста в NLP. Стемминг обеспечивает простое и быстрое уменьшение слов до их основ, тогда как лемматизация предоставляет более точные и контекстуально осведомленные результаты.

Стоп-слова

Стоп-слова — это слова, которые часто встречаются в тексте, но не несут значимой смысловой нагрузки для анализа. В контексте обработки естественного языка стоп-слова обычно удаляются из текста на этапе предобработки, чтобы улучшить эффективность и точность последующих операций, таких как классификация текста, извлечение информации и поиск.

Подходы к удалению стоп-слов:

1. Использование предопределенных списков: множество библиотек NLP предоставляют списки стоп-слов для различных языков. Например, библиотеки NLTK, SpaCy и sklearn содержат встроенные списки стоп-слов для английского и других языков.

2. Создание пользовательских списков: в зависимости от специфики задачи и корпуса данных, можно создать свои собственные списки стоп-слов, добавляя или исключая слова по мере необходимости.

3. Адаптивные методы: некоторые подходы предполагают автоматическое определение стоп-слов на основе их частотных характеристик в корпусе. Например, слова, которые встречаются очень часто или очень редко, могут быть помечены как стоп-слова.

Regex

Регулярные выражения (regex) — это последовательности символов, которые определяют шаблоны поиска в тексте. Регулярные выражения позволяют задавать сложные критерии поиска, включая поиск по шаблонам, замены и валидацию строк. Это мощный инструмент для работы с текстом в различных задачах, от простого поиска до сложной предобработки данных в NLP.

Bag-of-words

Метод "Bag of Words" (BoW) — один из наиболее простых и популярных методов представления текста в задачах обработки естественного языка. Этот метод преобразует текстовые данные в числовые векторы, что позволяет использовать их в моделях машинного обучения. Этот метод может применяться после того, как была произведена токенизация текста и сформирован словарь, состоящий из слов, наличие которых в тексте нас интересует.

Каждое предложение обрабатываемого текста в этом методе представляется в качестве вектора из нулей и единиц, имеющего длину, равную количеству слов в словаре интересующих нас слов, где единица говорит про то, что слово присутствует в предложении, а ноль – про его отсутствие [5].

Чем больший размер текста нужно обработать с помощью этого метода, тем больше станет словарь и размер вектора, который будет включать в себя много нулей. Чтобы избежать этого, можно уменьшить количество слов в словаре или создать более сложный словарь.

Метод BoW также можно расширить с помощью n-грамм для более детального учета контекста слов в тексте. N-граммы — это последовательности из n слов (или символов), которые могут предоставлять более богатую информацию о структуре и содержании текста по сравнению с отдельными словами. При использовании n-грамм словарь будет содержать не только отдельные слова, но и последовательности слов, что позволит учитывать более сложные зависимости и контексты.

При использовании n-грамм словарь будет содержать не только отдельные слова, но и последовательности слов, что позволит учитывать более сложные зависимости и контексты.

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) — это статистическая мера,

используемая для оценки значимости слова в документе по отношению к коллекции документов. Этот метод позволяет более точно учитывать важность слов в контексте, чем простой метод "Bag of Words" (BoW) [6].

Основные компоненты TF-IDF:

TF (Term Frequency): Частота слова в документе. Она измеряет, насколько часто слово встречается в данном документе, по формуле (1).

$$TF(t,d) = \frac{n_t}{\sum_k n_k}, \quad (1)$$

где n_t это число вхождений слова t в документ;

$\sum_k n_k$ – общее число слов в данном документе.

IDF (Inverse Document Frequency): Обратная частота документов. Она измеряет, насколько слово редкое или распространенное в коллекции документов, по формуле (2).

$$IDF(t,D) = \log\left(\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}\right), \quad (2)$$

где $|D|$ – число документов в коллекции;

$|\{d_i \in D \mid t \in d_i\}|$ – число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

TF-IDF: Произведение TF и IDF для слова. Оно показывает значимость слова в документе с учетом его распространенности в коллекции документов (формула 3).

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D). \quad (3)$$

Практическое применение методов NLP

Обработка естественного языка находит широкое применение в различных сферах, улучшая процессы и взаимодействие между людьми и машинами. Рассмотрим несколько ключевых областей применения методов NLP.

Анализ тональности

Анализ тональности (Sentiment Analysis) — это метод обработки естественного языка, направленный на определение эмоциональной окраски текста. Он позволяет выявлять позитивное, негативное или нейтральное отношение в текстовых данных. Анализ тональности широко применяется в бизнесе, маркетинге, социальных науках и многих других областях [7].

Приведем примеры использования анализа тональности.

Обслуживание клиентов: Анализ отзывов клиентов на продукцию или услуги для выявления общих настроений и удовлетворенности.

Социальные медиа: Мониторинг социальных сетей для оценки общественного мнения о бренде, продукте или событии.

Маркетинговые исследования: Анализ тональности отзывов и комментариев для понимания потребительских предпочтений и улучшения маркетинговых стратегий.

Политический анализ: Оценка общественного мнения о политических деятелях или инициативах на основе комментариев в социальных сетях и других источниках.

Методы анализа тональности

Лексиконные методы:

- основаны на предварительно созданных словарях, в которых каждому слову присваивается определенная эмоциональная оценка;

- примеры: SentiWordNet, AFINN, VADER (Valence Aware Dictionary and sEntiment Reasoner).

Машинное обучение:

- использование алгоритмов машинного обучения для классификации текста на основе тональности;

- примеры алгоритмов: Наивный Байес (Naive Bayes), Метод опорных векторов (SVM), Логистическая регрессия;

- требует предварительной разметки текстов для обучения модели.

Глубокое обучение:

- применение нейронных сетей для анализа тональности, что позволяет учитывать сложные зависимости и контексты в тексте;

П- римеры моделей: Рекуррентные нейронные сети (RNN), Долгосрочная краткосрочная память (LSTM), Трансформеры (BERT, GPT);

- модели глубокого обучения часто демонстрируют высокую точность, особенно при наличии большого объема данных для обучения.

Машинный перевод

Машинный перевод (Machine Translation, MT) — это процесс автоматического перевода текста с одного языка на другой с помощью компьютерных программ. Это одна из наиболее важных и широко используемых задач в области обработки естественного языка, которая имеет множество практических применений.

Принцип работы машинного перевода

Машинный перевод включает в себя несколько этапов:

1. Предварительная обработка текста: исходный текст подвергается предварительной обработке, которая включает в себя токенизацию (разделение текста на отдельные слова или фразы), удаление стоп-слов (например, предлогов, союзов) и нормализацию (приведение слов к нормальной форме).

2. Построение модели: для выполнения машинного перевода используются различные модели, от статистических методов до современных нейронных сетей. Эти модели

обучаются на больших корпусах параллельных текстов на разных языках.

3. Выравнивание: прежде чем перевести предложение на целевой язык, модель должна определить соответствие между словами и фразами исходного и целевого текста. Этот процесс называется выравниванием.

4. Перевод: используя полученные выравнивания и обученную модель, производится перевод исходного текста на целевой язык.

5. Постобработка: после перевода текст может быть подвергнут постобработке для улучшения качества перевода, например, для исправления грамматических ошибок или улучшения стиля перевода [8].

Применение машинного перевода

1. Коммерческий машинный перевод: В сфере бизнеса машинный перевод используется для перевода документов, писем, веб-сайтов и других текстовых материалов на разные языки для общения с клиентами и партнерами по всему миру.

2. Поисковые системы: Машинный перевод используется в поисковых системах для перевода результатов поиска или текстов рекламы на разные языки для максимизации охвата аудитории.

3. Социальные сети и коммуникация: В социальных сетях машинный перевод позволяет пользователям из разных стран общаться и взаимодействовать на их родном языке.

4. Автоматический переводчик: Машинный перевод используется в автоматических переводчиках для устного перевода разговоров и диалогов в реальном времени.

Технологии машинного перевода

1. Статистический машинный перевод: Основан на статистических моделях, которые анализируют большие объемы параллельных текстов на разных языках и извлекают статистические зависимости между ними.

2. Нейронные сети: Современные методы машинного перевода широко используют глубокое обучение и нейронные сети, такие как рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN) и трансформеры. Эти модели способны обрабатывать большие объемы текста и захватывать сложные зависимости между словами и фразами.

Автоматическое резюмирование

Автоматическое резюмирование — это процесс создания краткого и содержательного обзора или резюме из исходного текста. Этот процесс позволяет выделять ключевые и информативные аспекты текста, делая его более доступным и понятным для пользователей. Приведем некоторые особенности автоматического резюмирования (АР).

Принцип работы АР

1.Предварительная обработка текста: исходный текст подвергается предварительной обработке, которая может включать в себя токенизацию, удаление стоп-слов, лемматизацию и выделение предложений.

2.Выделение ключевых фраз и предложений: с помощью методов NLP выделяются ключевые фразы и предложения, которые наиболее информативны и репрезентативны для содержания текста.

3.Оценка важности: каждая фраза или предложение оценивается по степени их важности и информативности в тексте.

4. Генерация резюме: на основе оценок важности генерируется краткое резюме, включающее наиболее значимые и информативные аспекты исходного текста.

Применение АР

1. Сжатие текста: автоматическое резюмирование позволяет сокращать длинные тексты, делая их более компактными и удобными для чтения.

2. Информационные порталы: на информационных порталах автоматически создаются краткие обзоры новостей и статей, позволяя пользователям быстро получать основную информацию.

3. Анализ данных: в анализе данных автоматическое резюмирование помогает выделять ключевые аспекты и выводы из больших объемов текстовых данных.

4. Обработка документов: в бизнесе автоматическое резюмирование используется для обработки документов, отчетов и писем, помогая управлять информационными потоками.

Технологии АР

1. Извлечение ключевых слов: простые методы выделения ключевых слов на основе частотности и важности слов в тексте.

2. Извлечение ключевых предложений: анализ текста с использованием методов NLP для выделения ключевых предложений, содержащих наиболее важную информацию.

3. Генерация сжатого текста: использование алгоритмов для генерации кратких резюме на основе выделенных ключевых фраз и предложений.

4. Использование нейронных сетей: современные методы автоматического резюмирования используют глубокое обучение и нейронные сети для создания более точных и информативных резюме.

Чат-боты

Чат-боты — это программные агенты, способные автоматически взаимодействовать с

пользователями через текстовые интерфейсы, такие как мессенджеры, веб-сайты, приложения и другие. Они используются для решения различных задач, от предоставления информации и обработки заказов до выполнения операций в реальном времени.

Принцип работы:

1. Понимание запроса: чат-боты анализируют текстовые запросы пользователей и пытаются понять их смысл с помощью алгоритмов обработки естественного языка.

2. Генерация ответа: на основе понимания запроса чат-бот генерирует и отправляет ответ пользователю. Этот ответ может быть текстовым сообщением, ссылкой, изображением или даже действием.

3. Интерактивность: Некоторые чат-боты могут проводить диалог с пользователем, задавать дополнительные вопросы или запрашивать уточнения для более точного выполнения задачи.

Примеры применения:

Чат-боты используются в различных сферах, таких как клиентская поддержка, бронирование билетов, заказ товаров и услуг, проведение опросов, обучение и обучение, медицинская помощь и многое другое.

Заключение

Обработка текста методами естественного языка стала неотъемлемой частью современной технологии, играя ключевую роль в различных аспектах жизни и бизнеса. Эти методы позволяют компьютерам понимать, интерпретировать и генерировать человеческий язык, что открывает широкие возможности для автоматизации и улучшения коммуникации.

Анализ текста с использованием методов NLP позволяет автоматизировать множество процессов, включая анализ отзывов клиентов, мониторинг социальных сетей, улучшение поиска информации и поддержку многоязычного общения. Примеры практического применения этих методов демонстрируют их значимость в различных отраслях, таких как маркетинг, здравоохранение, образование и правоприменение.

Было рассмотрено несколько ключевых направлений NLP. Каждый из рассмотренных методов предоставляет уникальные инструменты для обработки и анализа текстовых данных, что делает их незаменимыми в разработке современных приложений.

Особое внимание в статье было уделено практическому применению методов NLP, в частности, в области анализа тональности, машинного перевода, автоматического резюмирования и чат-ботов. Эти технологии значительно улучшают взаимодействие между людьми и машинами, обеспечивая эффективную и естественную коммуникацию.

Несмотря на значительные достижения, NLP продолжает развиваться, сталкиваясь с новыми вызовами и открывая новые перспективы. Будущие исследования и разработки авторов будут направлены на улучшение точности и эффективности существующих методов, а также на создание новых подходов для решения более сложных задач. В этом контексте, NLP остается динамичной и важной областью, имеющей потенциал для значительного влияния на технологии и общество в целом.

Литература

1. Обработка естественного языка – [Электронный ресурс] / Интернет-ресурс. - Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Обработка_естественного_языка - Загл. с экрана (дата обращения: 24.05.24)

2. Правильный NLP: как работают и что умеют системы обработки естественного языка – [Электронный ресурс] / Интернет-ресурс. - Режим доступа: https://tproger.ru/articles/natural-language-processing/ - Загл. с экрана (дата обращения: 24.05.24)

3. Jurafsky, D., & Martin, J. H.. Speech and Language Processing (3rd ed.). - Pearson, 2021.

4. Thomas Landauer, Peter W. Foltz, & Darrell Laham. Introduction to Latent Semantic Analysis (англ.) // Discourse Processes (англ.) русск.: journal. – 1998. – Vol. 25. – Pp. 259–284. – DOI: 10.1080/01638539809545028.

5. Pennington, J., Socher, R., & Manning, C. D. Glove: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing *(EMNLP)*. – PP. 1532-1543.

6. Mark Needham. Scikit-learn: TF/IDF and cosine similarity for computer science papers. – 2017. – – [Электронный ресурс] / Интернет-ресурс. - Режим доступа: https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/ - Загл. с экрана (дата обращения: 24.05.24);

7. Pang, B., & Lee, L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, 2008. – T. 2(1-2). – P. 1-135.

8. Koehn, P. Statistical Machine Translation // Cambridge University Press, 2010.

Рудак Л.В., Зори С.А. Обработка текста методами естественного языка. *В работе рассматриваются методы обработки текста с использованием естественного языка*

(NLP), которые играют ключевую роль в современном мире информационных технологий. Статья охватывает основные концепции и техники NLP, такие как токенизация, стемминг, лемматизация, удаление стоп-слов, использование регулярных выражений, а также методы представления текста, включая Bag of Words и TF-IDF. Особое внимание уделено анализу тональности, машинному переводу, автоматическому резюмированию и чат-ботам, которые являются важными направлениями в области NLP.

Ключевые слова: *обработка естественного языка, машинное обучение, токенизация, искусственный интеллект, нейронные сети*

Rudak L.V., Zori S.A. Text processing using natural language methods. *The paper examines text processing methods using natural language (NLP), which plays a key role in the modern world of information technology. The article covers basic NLP concepts and techniques such as tokenization, stemming, lemmatization, stop word removal, regular expressions, and text representation methods including Bag of Words and TF-IDF. Particular attention is paid to sentiment analysis, machine translation, automatic summarization and chatbots, which are important areas in the field of NLP.*

Keywords: *natural language processing, machine learning, tokenization, artificial intelligence, neural networks*

Статья поступила в редакцию 20.06.2024
Рекомендована к публикации профессором Федяевым О. И.