Информационные технологии в обработке лингвистической информации

Мардона Шакировна Джумабаева
djumabayevamardona@gmail.com
Ринат Фаритович Бурнашев
rinat.burnashev@inbox.ru
Самаркандский государственный институт иностранных языков

Аннотация: В статье рассмотрены информационные технологии в обработке лингвистической информации. Более того показан процесс о принципах компьютерной лингвистики, методах извлечения и обработки информации, а также об основных алгоритмах и методах анализа языка.

Ключевые слова: обработка естественного языка, машинное обучение, анализ текста, машинный перевод, распознавание речи, генерация текста, компьютерная лингвистика, искусственный интеллект, многопоточность, большие данные, тотальность текста

Informational technologies in processing of linguistic information

Mardona Shakirovna Djumabayeva
djumabayevamardona@gmail.com
Rinat Faritovich Burnashev
rinat.burnashev@inbox.ru
Samarkand State Institute of Foreign Languages

Abstract: The article reveals the topic about the role of information technologies in the processing of linguistic information. Moreover, the information about the principles of computational linguistics is provided and methods of extracting and processing information, as well as the basic algorithms and methods of language analysis are distinctly discussed.

Keywords: Natural Language Processing (NLP), machine learning, text analysis, machine translation, speech recognition, text generation, computational linguistics, artificial intelligence, multithreading, huge data, text totality

С развитием современных технологий всё больше и больше областей человеческой деятельности становятся связанными с компьютерами. Одной из

таких областей является лингвистика. Компьютерная лингвистика - это наука, которая занимается автоматической обработкой языка с помощью компьютерных технологий. Она является частью искусственного интеллекта и включает в себя такие направления как машинный перевод, распознавание речи, анализ текстов и многое другое. Компьютерная лингвистика позволяет решать множество задач, связанных с языком, что делает её незаменимой в современном мире. В данной статье мы рассмотрим главные принципы работы компьютерной лингвистики и её основные приложения.

Обработка естественного языка (Natural language processing, NLP) - это область, посвященная разработке алгоритмов и компьютерных программ, которые позволяют компьютерам «понимать» естественные языки, такие как английский, немецкий, французский и т.д. Они могут использоваться для создания различных систем интеллектуального анализа текста, включая машинный перевод, анализ тональности и сентиментов, автоматическую классификация текстов и многое другое.

Обработка естественного языка использует различные методы для анализа естественного языка, включая статистический анализ, машинное обучение и глубокое обучение. Он имеет множество приложений в различных областях, таких как медицина, право, образование, реклама, финансы и многие другие. Задачи, решаемые при помощи данной технологии, включают в себя часто-речной анализ тональности и эмоций, а также разбор сложных синтаксических конструкций. Обработка естественного языка является технологической основой для создания множества современных приложений, таких как голосовые помощники, персонализированные рекомендации, системы автоматической обработки текстов, и многие другие.

Основные задачи обработки естественного языка включают:

1. Разбор текста на отдельные слова и выделение их формы и значения (морфологический анализ). Это технология, которая используется для анализа структуры слов в тексте (например, определение основы слова, окончания и т.д.), что может быть полезно для анализа грамматических ошибок, автоматического исправления отпечаток и других задач связанных с обработкой естественного языка.

2. Определение грамматической структуры предложения и взаимосвязей между его частями (синтаксический анализ). Синтаксический анализ - это одна из задач обработки естественного языка, которая заключается в анализе грамматической структуры предложения и определении взаимосвязей между его частями.

Существуют два основных типа анализа: морфологический анализ и синтаксический анализ. Морфологический анализ заключается в разделении

слова на его составляющие (морфемы) и определения их формы и значения. Синтаксический анализ определяет структуру предложения и взаимосвязи между его частями.

Синтаксический анализ может быть проведен вручную, когда человек просматривает текст и анализирует его грамматическую структуру. Однако, такой подход является очень трудоёмким и неэффективным при работе с большими объёмами текста. Поэтому синтаксический анализ часто проводится автоматически при помощи компьютерных программ.

Существуют различные методы автоматического синтаксического анализа. Некоторые из них основаны на правилах и эвристиках, т.е. заранее заданных грамматических правилах. Другие методы используют машинное обучение для создания моделей, которые могут распознавать грамматические структуры предложения. Кроме того, некоторые методы используют комбинацию правил и машинного обучения.

Синтаксический анализ является важным шагом в обработке естественного языка и может быть использован для решения различных задач, таких как машинный перевод, извлечение информации из текста, классификация текстов и многое другое.

3. Анализ смысла слов и выражений в контексте текста (семантический анализ). Семантический анализ - это анализ значения слова и выражения, то есть изучение того, какое значение они несут в конкретном контексте. Этот анализ может проводиться на уровне отдельных слов, фраз и предложений. Например, в предложении «Я купил новый автомобиль», семантический анализ позволяет определить значение слов «новый» и «автомобиль» в данном контексте.

4. Извлечение именованных сущностей, таких как имена людей, организаций и географических мест (Named Entity Recognition, NER) - это задача обработки естественного язык4а, которая заключается в определении именованных сущностей в тексте, таких как имена личностей, компаний, организаций, географических мест и т.д. NER является важной задачей обработки естественного языка, поскольку именованные сущности могут иметь большое значение для понимания содержания текста. Например, в новостных статьях имена людей и названия организаций могут указывать на ключевые фигуры и события. В финансовых или бизнес-новостях, имена компаний и географические места могут указывать на конкретные места и события.

Существуют множество методов для проведения NER, включая правила и эвристики, статистические модели и глубокое обучение. Более простые методы могут использовать заранее заданный словарь именованных сущностей и правила для определения их вхождения в текст. Более сложные методы могут

использовать машинное обучение для создания моделей, которые могут распознавать именованные сущности в тексте.

NER может использоваться для решения различных задач, таких как извлечение информации из текста, автоматическая классификация документов, машинный перевод, анализ настроений и многое другое.

5. Определение тональности и эмоциональной окраски текста (анализ тональности). Анализ тональности - это процесс определения эмоциональной окраски текста, который может быть положительным, отрицательным или нейтральным.

Анализ тональности может проводиться на уровне отдельных слов, фраз и предложений, а также на уровне целого текста. Например, в предложении «Я очень доволен своим новым автомобилем», анализ тональности позволяет определить, что это положительное высказывание с использованием слова «доволен».

Анализ тональности широко используется в маркетинге, социальных исследованиях и других областях для изучения мнений и настроений людей.

6. Машинный перевод с одного языка на другой. Машинный перевод - это задача обработки естественного языка, которая заключается в автоматическом переводе текста с одного языка на другой язык.

Существуют множество методов машинного перевода, но основными подходами являются:

- *статистический машинный перевод*: основан на идеи, что перевод одного предложения с одного языка на другой язык должен быть похож на перевод других предложений с аналогичной грамматической структурой. Он использует большие объёмы параллельных корпусов текста, чтобы вычислить вероятности перевода различных слов и фраз;
- *машинное обучение*: использует нейронные сети для создания моделей машинного перевода на основе больших объёмов параллельных текстов;
- *глубокий машинный перевод*: это подход к машинному обучению, который использует глубокие нейронные сети для перевода текстов.

Машинный перевод может быть использован для автоматического перевода больших объёмов текста, что может быть полезно для деловых целей, межкультурных коммуникаций и многих других сфер. Однако, машинный перевод не всегда обеспечивает достаточно точный и чёткий перевод, особенно для контента, который имеет сильную эмоциональную окраску или использует многозначные слова и выражения.

7. Генерация текста на основе заданных параметров и условий (текстогенерация). Текстогенерация - это процесс создания текста компьютером или программой. В зависимости от задачи, тексты могут быть сгенерированы как

на основе готовых данных, так и на основе алгоритмов искусственного интеллекта. Например, в автоматической рекламе можно использовать текстогеренацию для создания уникальных объявлений, а в журналистике - для написания статей на основе данных и фактов.

Текстогенерация также может использоваться для создания синтетических текстов, таких как поэзия или литературные произведения. Однако, хотя компьютеры могут генерировать тексты, они не могут понимать смысл и контекст слов и фраз, поэтому качество текста может быть недостаточным для некоторых задач.

8. Классификация текстов по заданному критерию, например, по тематике или жанру. Классификация текстов - это задача обработки естественного языка, которая заключается в автоматической классификации текстов на основе заданного критерия, например, по тематике, жанру, настроению и т.д.

Для классификации текстов используются различные методы, включая методы машинного обучения, статистические методы и комбинированные методы. Они используются для распознания особенностей текста в зависимости от заданного критерия, формируются различные признаки, такие как частота слов, структура предложений, использование ключевых слов и так далее. Затем алгоритмы классификации обучаются на основе этих признаков.

Некоторые из методов классификации текстов включают:

- *метод наивного Байеса*: использует статистические методы на основе частотности слов и фраз в текстах;
- *метод максимальной энтропии*: использует теорию информации для поиска наилучшей модели классификации текстов на основе их свойств;
- *метод опорных векторов*: использует машинное обучение для построения модели, которая может быть разделять тексты на различные классы.

Методы классификации текстов могут быть использованы для автоматической категоризации больших объёмов текста, что может быть полезно для сортировки информации, анализа социальных сетей и многих других задач.

9. Распознавание голоса и естественной речи. Распознавание голоса и естественной речи - это задача обработки естественного языка, которая заключается в переводе устной речи в текстовую форму и обратно.

Распознавание голоса и естественной речи состоит из двух основных этапов:

- *распознавание речи*: процесс, который заключается в преобразовании звуков человеческой речи в признаки, которые могут быть интерпретированы компьютерной программой;

• *обработка естественного языка*: процесс, который заключается в интерпретации текста, полученного от распознавания речи.

Существуют различные методы для проведения распознавания голоса и естественной речи, включая использование скрытых марковских моделей, нейронных сетей, глубокого обучения и комбинации этих методов.

Распознавание голоса и естественной речи может быть использовано для многих приложений, включая системы распознавания голосовой идентификации, системы автоматического анализа звуков, системы перевода речи и многое использован для облегчения доступа к информации для людей, с ограниченными возможностями., такими как слабовидящие и слабослышащие.

Задачи обработки естественного языка могут применяться в различных сферах, таких как медицина, право, финансы, маркетинг, образование и многие другие.

Методы и подходы к обработке естественного языка включают:

1. Статистические методы: используют математические методы для анализа больших объёмов текста и выявления статистических закономерностей. Они широко используются в компьютерной лингвистике, машинном обучении, обработке естественного языка и других областях, где требуется анализ больших массивов данных.

Некоторые из наиболее распространённых статистических методов включают в себя:

• *частотный анализ*: определение частоты встречаемости слов или фраз в тексте для определения ключевых слов или выражений;

• *кластерный анализ*: группировка текстовых документов на основе их схожести;

• *методы машинного обучения*: использование алгоритмов машинного обучения для классификации текстовых документов иди извлечения информации;

• *анализ синтаксических отношений*: определение синтаксических отношений между словами и фразами в тексте;

• *анализ семантических связей*: определение семантических связей между словами и фразами в тексте для определения их значения и контекста.

Статистические методы позволяют сделать выводы на основе данных и установить закономерности, которые могут быть использованы для принятия решений в различных областях.

2. Машинное обучение: использует алгоритмы машинного обучения для построения моделей, которые могут распознавать и интерпретировать языковые данные. Это позволяет компьютеру анализировать большие объёмы текстов и находить закономерности в них, которые можно использовать для создания

новых текстов. Например, алгоритмы машинного обучения могут использоваться для создания персонализированных рекомендаций на основе интересов пользователя или для автоматического перевода текстов на другой язык. Однако, хотя машинное обучение и текстогенерация имеют большой потенциал, они также могут столкнуться с рядом проблем. Например, компьютер может неправильно понимать смысл слов и фраз, что может привести к созданию некорректных или несвязанных текстов. Кроме того, использование текстогенерации может подвергнуться критике за отсутствие оригинальности и творческого подхода. Тем не менее, современные алгоритмы машинного обучения и текстогенерации продолжают развиваться, и в будущем они могут стать ещё более точными и эффективными. Это может привести к созданию новых инструментов для автоматизации процесса написания текстов и улучшения качества текстовых данных в целом.

3. Глубокое обучение: это подход к машинному обучению, который использует нейронные сети для обработки текстов и извлечения информации. Глубокое обучение - это подход к машинному обучению, использующий многослойные нейронные сети для обработки данных и извлечения признаков (в том числе текстовых), а также для принятия решений на основе полученных данных.

В контексте обработки текстов, глубокое обучение может использоваться для различных задач, таких как:

- *классификация текстов*: в определение категории, в которую можно отнести заданный текст (например, определение жанра книги по её описанию или тематики письма по его содержанию);
- *извлечение информации*: автоматическое извлечение фактов и данных из текстовых источников (например, перечисление всех имён и фамилий в новостной статье);
- *машинный перевод*: перевод текстовых документов с одного языка на другой с помощью нейронных сетей;
- *генерация текста*: автоматическое создание текста на основе заданных правил и структур (например, создание описания товара или генерация синтетической речи). Глубокое обучение является одной из наиболее перспективных областей искусственного интеллекта и находит применение в различных задачах, связанных с обработкой текстов и естественного языка.

4. Правила и эвристики: используют заранее заданные правила и эвристики для анализа и интерпретации текстов. Эти правила могут быть связаны с грамматикой, семантикой и контекстом текста. Например, такие правила могут использоваться для автоматической классификации текстов по тематике или для определения тональности текста. Однако, такой подход имеет свои ограничения,

поскольку он может быть применён только к текстам, которые соответствуют заранее заданным правилам. Кроме того, такой подход может быть менее точным в случае текстов с нестандартной грамматикой или необычным контекстом. В целом, использование правил и эвристик может быть полезным инструментом для анализа текстов, особенно если они используются в сочетании с другими методами анализа, такими как машинное обучение и текстогенерация.

5. Комбинированные методы: сочетают в себе различные методы, чтобы достичь наилучшей производительности и точности обработки текстов. Комбинированные методы в обработке текстов используют несколько различных алгоритмов и подходов, чтобы достичь наилучшей производительности и точности. Например, комбинированные методы могут включать в себя следующее:

- *использование статистических методов для извлечения ключевых слов или фраз*, а затем использование методов глубокого обучения для классификации текстов с учётом этих ключевых слов.
- *использование классических методов машинного обучения*, таких как наивный байесовский классификатор, а затем дополнительно обучение многослойной нейронной сети на выходах этого классификатора для повышения точности и производительности.
- *использование технологий обработки естественного языка* (NLP), таких как разметка и лемматизация, совместно с методами машинного обучения и глубокого обучения для улучшения анализа текста.

Комбинированные методы могут быть эффективными, когда использование только одного метода не даёт достаточной точности или производительности в обработке текста. Поэтому они могут стать очень полезными в таких областях, как анализ социальных медиа, мониторинг новостей, автоматизированное обслуживание клиентов и т.д.

6. Обработка естественного языка на основе знаний: использует базы знаний и онтологии для анализа и интерпретации текстов. Этот подход позволяет более точно понимать смысл текстов, так как базы знаний содержат информацию о связях между словами, понятиями и фактами. Он может быть использован для автоматического извлечения информации из текстов, ответа на вопросы и построения диалоговых систем.

Однако, этот подход также имеет свои ограничения, так как он требует больших затрат на создание и поддержание баз знаний и онтологий. Кроме того, он может быть менее эффективным в случае нестандартных текстов или новых областей знаний, для которых не были созданы соответствующие базы знаний. В целом, обработка естественного языка на основе знаний может быть эффективным инструментом для анализа текстов, особенно в областях, где есть

хорошо разработанные базы знаний и онтологии. Однако, для более широкого применения этот подход требует дополнительных исследований и разработок.

7. Многопроходная обработка: использует несколько проходов по тексту для обнаружения и анализа различных аспектов языка.

Многопроходная обработка - это метод обработки текста, в котором текст проходит через несколько этапов или проходов, каждый из которых отвечает за обработку разных аспектов языка. Например:

- *синтаксический анализ*: первый проход осуществляет синтаксический анализ текста для выделения частей речи, определения связей между словами, построения дерева зависимостей и т.д.
- *семантический анализ:* на втором проходе производится анализ семантических связей между словами, определение значения терминов, выделение понятий и т.д.
- *определение тональности*: на третьем проходе текст анализируется на наличие положительных/отрицательных выражений, тональных слов и т.д.
- *извлечение информации*: на последнем проходе происходит извлечение необходимой информации из текста, такой как факты, даты, имена людей и организаций и т.д.

Преимуществом многопроходной обработки является более глубокий анализ текста и более точное определение его смысла, благодаря тому, что каждый проход посвящен определенным аспектам текста. Недостатком этого метода является то, что он требует больше ресурсов и времени на обработку текста, поэтому может быть не очень эффективным для быстрого анализа больших объемов данных.

Современные технологии обработки естественного языка (Natural Language Processing, NLP) включают в себя вышеперечисленные методы и алгоритмы, которые позволяют компьютерам обрабатывать и анализировать естественный язык, используемый людьми для общения.

Все эти технологии и методы используются в различных приложениях и системах, таких как *чат-боты*, *анализаторы текста*, *системы автоматического ответа на электронные письма* и многое другое.

Обработка естественного языка - одно из наиболее интересных и перспективных полей исследований в области компьютерной лингвистики, искусственного интеллекта и машинного обучения. Эта область охватывает множество задач, связанных с обработкой и анализом естественного языка, машинным переводом, распознаванием речи и прочие.

С появлением алгоритмов обучения и разработкой сложных моделей, способных анализировать тексты, задачи обработки естественного языка стали

решаться намного эффективнее. Однако, несмотря на значительный прогресс в этой области, ещё многое можно улучшить и усовершенствовать.

Обработка естественного языка имеет многочисленные практические применения в различных областях, таких как банковское дело, медицина, научные исследования, лингвистика, анализ графических данных и др. Это позволяет использовать эту область для повышения эффективности работы в разных сферах.

В целом, обработка естественного языка является перспективной областью, которая непрерывно развивается и находится в центре внимания многих исследователей и разработчиков программного обеспечения. Её применение всё больше популяризируется в повседневной жизни, делая её важнейшей областью развития искусственного интеллекта.

Использованная литература

1. Бурнашев Р. Ф., Мустафина А. Д. Синтаксический анализ как инструментарий квантитативной лингвистики //Science and Education. - 2022. - Т. 3. - №. 12. - С. 1211-1220.

2. Бурнашев Р. Ф., Аламова А. С. Роль нейронных сетей в лингвистических исследованиях //Science and Education. - 2023. - Т. 4. - №. 3. - С. 258-269.

3. Бурнашев Р.Ф. и др. Роль экспертных систем в лингвистических исследованиях //Science and Education. - 2023. - Т. 4. - №. 3. - С. 941-950.

4. Аламова А. С., Бурнашев Р. Ф. Контент-анализ как инструментарий квантитативной лингвистики при изучении художественных текстов //Science and Education. - 2022. - Т. 3. - №. 12. - С. 1188-1200.

5. Бурнашев Р.Ф., Аламова А.С. Квантитативная лингвистика и искусственный интеллект //Science and Education. - 2022. - Т. 3. - №. 11. - С. 1390-1402.

6. Бурнашев Р.Ф., Ахророва Ф.Р. Роль информационных технологий в определении частотных характеристик объектов //Science and Education. - 2022. - Т. 3. - №. 11. - С. 571-582.

7. Бурнашев Р. Ф., Фаррухова Ф. Ш. Лингвистический корпус как база для организации информационного поиска //Science and Education. - 2021. - Т. 2. - №. 3.

8. Бурнашев Р. Ф., Мирзаева А. Б. Контент-анализ как инструментарий квантитативной лингвистики //Science and Education. - 2022. - Т. 3. - №. 12. - С. 1201-1210.

9. Насырова Г. Н., Амонова Ш. Х., Бурнашев Р. Ф. Обзор современных сервисов и программного обеспечения квантитативной лингвистики //Science and Education. - 2022. - Т. 3. - №. 12. - С. 450-462.

10. Мансур Ж. Д. Н. З., Саттарова А. Т., Бурнашев Р. Ф. Роль лингвистических корпусов в создании и совершенствовании систем машинного перевода //Science and Education. - 2022. - Т. 3. - №. 2. - С. 1348-1358.

11. Бурнашев Р. Ф., Ахадова Ш. С., Нематуллаева Н. Б. К вопросу об особенностях лингвистических корпусов второго и третьего поколений //ЕВРОПА, НАУКА И МЫ: сборник научных публикаций международной научно-практической конференции.-Издательство «Education and Science» Чехия, Прага. - 2021. - С. 77-79.

12. Бурнашев Р. Ф., Болтаева Н. С., Абилова К. М. Применение лингвистических корпусов для определения сложности текста //ЕВРОПА, НАУКА И МЫ: сборник научных публикаций международной научно-практической конференции.-Издательство «Education and Science» Чехия, Прага. - 2021. - С. 79-82.

13. Бурнашев Р. Ф., Нематуллаева Н. Б., Худоярова П. Н. Роль лингвистических корпусов в научных исследованиях //SCIENCE AND EDUCATION: сборник научных публикаций международной научно-практической конференции. - Турция, Анталия. - 2021. - С. 126-128.

14. Бурнашев Р. Ф., Фаррухова Ф. Ш. Особенности использования облачных технологий в современных условиях //Science and Education. - 2021. - Т. 2. - №. 3. - С. 200-205.