Джангаров А.И.¹, Калхиташвили Д.Ш.², Магамедова Д.М.³

Машинная обработка текста

¹Чеченский государственный университет имени А.А.Кадырова
(Россия, Грозный)

²Российская академия народного хозяйства и государственной службы при Президенте
Российской Федерации
(Россия, Москва)

³Чеченский государственный педагогический университет
(Россия, Грозный)

doi: 10.18411/trnio-12-2021-78

Аннотация

Машинная обработка текста является очень важным инструментом в современной методологии обработки данных. В настоящее время машинная обработка естественного языка позволяет выстроить современную и удобную форму коммуникации между человеком и машиной. Благодаря методам формализации текста, становится возможным применение NLP, например, в области диалоговых систем, персонализированных ботов-помощников, систем информационного поиска, генерации текстового контента и других прикладных областях. В данной исследовательской работе освещаются основные положения и методы машинной обработки текста.

Ключевые слова: машинная обработка, обработка данных, большие данные, NLP.

Abstract

Machine word processing is a very important tool in modern data processing methodology. Nowadays, machine processing of natural language makes it possible to build a modern and convenient form of communication between man and machine. Thanks to the text formalization methods, it becomes possible to apply NLP, for example, in the field of dialogue systems, personalized helper bots, information retrieval systems, text content generation and other application areas. This research paper highlights the main provisions and methods of machine text processing.

Keywords: machine processing, data processing, big data, NLP.

На сегодняшний день имеется несколько специализированных систем, занимающихся обработкой неструктурированной текстовой информации. Зачастую, с первого взгляда довольно сложно разобраться во всех типах обработки. Именно поэтому необходимо выделять системность в таком разнообразии подходов к анализу обработки текстов. Помимо удобства восприятия, это облегчит выбор алгоритмов для каждой задачи.

В качестве текстов для анализа может выступать различная информация из таких сфер деятельности как, например, научные исследования, медицина, юридическая практика и так далее. Целевые задачи для каждой сферы могут различаться, но есть некоторый свод задач, которые будут общими для всех сфер. К таким задачам можно отнести, например, проверку орфографии, машинный перевод отдельных слов. Сейчас такие задачи не вызывают затруднений в их формализации. Тем не менее, сейчас становятся всё более интересны и другие задачи, связанные с конкретной спецификой области. Так, например, всплывающие подсказки слов при наборе текста должны иметь смысловой контекст в рамках терминологии определенной сферы деятельности. Кроме того, сейчас актуальны задачи вопросно-ответных систем, сентиментного анализа, машинного перевода. Для решения такого рода задач требуются намного более большие корпуса текстов. Требование к большому объему данных обусловлено спецификой работы таких алгоритмов.

Вышеуказанные проблемы достаточно непросто формализовать. Они не имеют не только единственно верного решения, но и заранее определенного множества допустимых

решений. Все решения должны рассматриваться в контексте поставленной задачи. Такие решения достаточно сложно оценивать. Есть подходы, которые позволяют упростить анализ результатов. Среди таких подходов можно выделить составление ключевых терминов в тексте (коллокаций), суммаризация (реферирование) текста. Также могут быть полезны и методы визуально-графического представления результатов исследований.

Многие современные методы обработки естественного языка берут свое математическое начало в компьютерной лингвистике, в частности, в сетях Маркова. Классические алгоритмы подвергаются изменениям для того, чтобы отвечать требованиям к обработке больших объемов данных (Big Data).

Классические алгоритмы адаптируются к современным объемам данных благодаря нескольким подходам. Одним из таких подходов является организация параллельной работы алгоритмов. Кроме того, возможно и уменьшение размерности данных с помощью таких методов, как, например, PCA (principal component analysis – метод главных компонент). Также при обработке больших данных – как, впрочем, и других – необходим препроцессинг данных, который включает в себя этапы парсинга и нормализации. Такие подходы применимы для многих языков, включая русский, в котором существует большое количество различных словоформ.

Компьютерная лингвистика берет свое историческое начало с 1954 года, когда в Джорджтаунском университете (Нью-Йорк, США) был проведен эксперимент по машинному переводу более 60 предложений с русского на английский.

Далее область обработки естественных языков получила свое следующее развитие: в 60-х годах прошлого столетия начался период ещё большей популярности идей искусственного интеллекта. Именно тогда была создана компьютерная программа, которая была призвана проводить психологическую терапию. Программа была задумана для диалога с пользователем, в ходе которого она формировала свои ответы с целью создать связный текст, который был бы максимально похожим на привычный язык общения людей.

Между периодами бурного роста область обработки естественных языков переживала и затишья. Так, например, оно продолжалось вплоть до 80-х. Именно тогда начала активно развиваться идея корпусной лингвистики. Эта идея заключалась в сборе, обработке и использовании больших объемов текста (корпусов) из различных тематических направлений (предметных областей). В то время уже появились компьютеры, которые были способны полностью хранить эти данные, обрабатывать их и вычислять необходимые значения по таким корпусам. В основном, расчет значений был в области статистических данных по отдельным токенам (словам). Все это было необходимо для того, чтобы строить модели машинного обучения, которые обрабатывали текстовые данные.

Во время 90-х область статистической обработки текста продолжила свое планомерное развитие. Так, за эти годы произошло более полное углубление в анализ текстовых данных посредством статистической обработки корпуса текстов.

Сейчас состояние области анализа естественных языков бурно развивается. Особенно это заметно на примере английского языка, который, де-факто, сейчас является языком международного общения. В русском же языке не так хорошо изучены и формализованы все нюансы языка. В основном сейчас идет адаптация методов обработки английского языка для русского. Уникальные разработки есть, но их не так много, как хотелось бы. Тем не менее, данная область активно развивается.

Системы автоматизированной обработки текста могут быть разделяться на несколько типов. Одним из таких факторов деления является лицензия на программный продукт. Бывают как открытые системы (open source), которые могут быть свободно распространены и модифицированы другими разработчиками, так и закрытые. Закрытые (проприетарные) являются чьей-то собственностью (например, собственностью организации).

Также системы автоматизированной обработки текста принято делить по принципу принадлежности к языку. Бывают системы для определенного языка, а бывают многоязычные. Многоязычные системы обладают большей универсальностью, но зачастую отстают по качеству от специализированных систем для конкретного языка. Тем не менее

их универсальность позволяет снизить цену коммерческой эксплуатации за счет наработок в этой области со всего мира.

Одним из таких примеров многоязычных систем является система автоматического перевода «Google Translate». Данная система использует искусственный «универсальный» язык для перевода. Поэтому данная система способна переводить текст напрямую с одного естественного языка на другой. Другие подобные системы ранее использовали английский язык в качестве промежуточного, из-за чего качество перевода нередко снижалось.

Как уже было сказано ранее, идея создания корпусов языка зародилась еще в 80-х годах. Тем не менее, эта идея актуальна и по сей день. Корпусы текста являются одной из главных частей для многих алгоритмов обработки естественных языков.

Таким образом, на сегодняшний день машинная обработка текстов играет большую роль в жизнедеятельности человека и знания в данной области оцениваются крайне высоко. На рынке существует дефицит квалифицированных специалистов данной области и можно с уверенностью утверждать то, что данное направление будет актуально всегда и является отличным выбором для изучения и освоения практических инструментов обработки текстов.

1. Как «КиберЛенинка» вывела Россию в лидеры открытого доступа к научным работам [Электронный ресурс]. – URL: <https://te-st.ru/2015/07/30/kiberleninka> (дата обращения 08 октября 2019)
2. Йоаким Нивре, Игорь М. Богуславский, Леонид Л. Иомдин, 2008 «Анализ деревьев синтаксиса русского языка» [Электронный ресурс]. - URL: <https://www.aclweb.org/anthology/C08-1081.pdf> (по состоянию на 17 октября 2019 г.). Суханов А.А. Маратканов А.С. Анализ способов хранения социальных данных из сети Интернет. //International scientific review. Москва, 2017.
3. Руководство морфологического анализатора pymorphy2 [Электронный ресурс]. – URL: <https://pymorphy2.readthedocs.io/en/0.2/user/index.html> (дата обращения 06 сентября 2019).

Джангаров А.И.¹, Калхиташвили Д.Ш.², Магамедова Д.М.³

Обзор экосистемы мессенджеров

¹Чеченский государственный университет имени А.А.Кадырова
(Россия, Грозный)

²Российская академия народного хозяйства и государственной службы при Президенте
Российской Федерации
(Россия, Москва)

³Чеченский государственный педагогический университет
(Россия, Грозный)

doi: 10.18411/trnio-12-2021-79

Аннотация

На начало 2021 года количество интернет-пользователей выросло на 298 миллионов, по сравнению с прошлым годом. Также и выросла потребность в информации. Пользователи проводят все больше времени в интернете, решая там все больше задач. В связи с пандемией, бизнес, мероприятия и обучение переходят на онлайн-формат, так как это самый доступный и безопасный формат получения услуг и информации. Кроме источника информации интернет может быть полезен и в общении. Связь с коллегами, друзьями и семьей, новые знакомства – все это теперь является элементами повседневной жизни, причем выбор места для онлайн-общения достаточно велик. В данной работе приводится обзор и особенности экосистем популярных мессенджеров.

Ключевые слова: мессенджер, WhatsApp, Facebook, Telegram, чат-бот, онлайн, экосистема, интернет.

Abstract

At the beginning of 2021, the number of Internet users grew by 298 million compared to last year. The need for information has also grown. Users spend more and more time on the Internet, solving more and more problems there. Due to the pandemic, business, events and training are