STUD NET

NLP - ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ

NATURAL LANGUAGE PROCESSING

УДК-004

Цитульский Антон Максимович, студент, МГТУ им. Н. Э. Баумана, Россия, г. Москва
Иванников Александр Владимирович, студент, МГТУ им. Н. Э. Баумана, Россия, г. Москва
Рогов Илья Сергеевич, студент, МГТУ им. Н. Э. Баумана, Россия, г. Москва
Tsitulsky Anton Maksimovich, tmath90@yandex.ru
Ivannikov Alexander Vladimirovich, tmath90@yandex.ru
Rogov Ilya Sergeevich

Аннотация

Статья посвящена исследованию сущности обработки естественных языков в рамках перспектив развития машинного обучения. Автором отмечено, что обработка естественного языка активно исследуется в наше время. Также выделено, что существует много методов, с помощью которых можно преодолеть трудности в работе систем обработки естественного языка. На практике применение таких методов требует дальнейшего улучшения и совершенствования посредством глубокого машинного обучения.

Abstract

The article is devoted to the study of the essence of processing natural languages in the framework of the prospects for the development of machine learning. The author noted that natural language processing is actively studied in our time. It is also highlighted that there are many methods with which you can overcome the difficulties in the operation of natural language processing systems. In practice, the application of such methods requires further improvement and improvement through deep machine learning.

Ключевые слова: естественные языки, обработка, машинное обучение, глубоко обучение, нейронная сеть.

Keywords: natural languages, processing, machine learning, deep learning, neural network.

За два последних века человечество успешно справилось с автоматизацией многих задач используя механические и электрические приборы. Вторая половина XX века обратила внимание человека к автоматизации обработки

естественного языка. Теперь человеку нужна помощь не только с механической работой, но и с интеллектуальными задачами. Задачей для машин в современных условиях ставится способность читать неподготовленный текст, проверять его на ошибки, выполнять задачи, поставленные в тексте, и даже понимать текст настолько хорошо, чтобы дать в ответ, основываясь на значении текста. ***Актуальность*** данной статьи обуславливает наличие проблем обработки естественного языка, котрую нельзя назвать простой. Трудности возникают по ряду объективных причин, среди которых существование сотен естественных языков, имеющих свои синтаксические правила. Кроме того, в рамках одного языка существуют слова, которые могут иметь смешанное содержание в зависимости от контекста употребления. Даже на уровне отдельных символов встречаются определенные трудности. ***Целью*** статьи является исследование современных процессов обработки естественных языков (NLP).

В процессе обработки естественного языка всегда следует учитывать кодирование, используемое в конкретном документе. Текст может храниться в разных кодировках: ASCII, Unicode, UTF-8, UTF-16, Latin-1 и др. Особые виды обработки могут понадобиться для знаков пунктуации и для чисел. Также иногда приходится отдельно обрабатывать использование знаков, которые отражают эмоции (комбинации символов или специальные символы), гиперссылок, повторяющихся знаков препинания (... или ---), расширений файлов и имен пользователей, содержащих точки [3].

Под распределением текста на фрагменты или элементы обычно имеется в виду представление текста в виде последовательности слов. В этом случае слова обозначаются термином «лексический элемент», «лексема», или просто «токен» (token), а процесс разделения текста - «токенизация» (tokenization). Этот процесс не представляет трудности в языках, использующих символы-пробелы для разделения слов, но в языках, подобных китайскому, это сделать гораздо труднее, поскольку иероглифы могут обозначать как склады, так и целые слова. Также и в английском языке с процессом токенизации могут возникнуть определенные трудности, ведь существует большое количество альтернативных вариантов, когда одно слово может писаться слитно, раздельно или через дефис [2].

Слова объединяются в словосочетания и предложения, следовательно, определение границ предложений тоже может быть связано с определенными трудностями, хотя на первый взгляд кажется, что достаточно только найти точку, обозначающую конец предложения. Но точки могут встречаться и внутри предложений, например, после сокращенных слов [3]. При

грамматическом разборе все еще возникают серьезные проблемы с точностью. Во-первых, многое здесь зависит от качества морфологической (части речи) разметки (part-of-speech tagging), которая должна быть очень высокой (97-98%), однако в длинных предложениях очень часто можно встретить неправильно распознанную определенную часть речи, что приводит к дальнейшим ошибкам разбора. Во-вторых, современный автоматический синтаксический разбор дает точность примерно 90-93%, а это, в свою очередь, означает, что в длинном предложении практически всегда будут ошибки разбора. Например, при точности разбора 90%, вероятность разбора предложения длиной 10 слов без единой ошибки составит всего 35%.

Часто правильный синтаксический разбор включает также понимание семантики предложения, но, например, в английском языке это нередко вызывает трудности. Так, в предложении «He saw a man with a hammer» может быть два варианта синтаксического разбора в зависимости от того, считаем ли мы, что человека увидели с помощью молотка или увидели человека с молотком. Конечно, если нужно получить максимально точный синтаксический разбор, то имеет смысл оставлять несколько наиболее вероятных вариантов, а затем определять правильный по совокупности различных факторов, в том числе семантических [1].

Иногда приходится определять связи между словами. Например, установление кореферентности (coreference resolution) определяет связи между конкретными словами, обозначающими один и тот же объект, то есть имеющими один и тот же референт в одном или в нескольких предложениях. Например, в предложениях «The city is large but beautiful. It fills the entire valley» слово «it» кореферирует, то есть референционно тождественному слову «city».

Явления кореферентности обусловлены фундаментальными закономерностями организации текста. Поскольку текст имеет линейное строение, а ситуация, которую он описывает, как правило, нелинейная, в тексте почти неизбежно должны содержаться повторные упоминания элементов ситуации, которая описывается. При каждом новом упоминании того же объекта проводится новая номинация этого объекта, основанная на том, что уже было сказано об этом объекте, и на тех знаниях, которые в тексте не вербализированы (экстралингвистические знания говорящего о контексте предметной области).

Хотя проблема кореферентности в лингвистике достаточно подробно исследована, воплощение этих теоретических знаний на практике на сегодняшний день является достаточно сложным [4]. Если слово может иметь

несколько смысловых значений, для определения его смысла в данном конкретном случае может потребоваться выполнение операции решения лексической многозначности (word sense disambiguation, WSD). Это связано с определенными трудностями. Например, в предложении «John went back home» слово «home» может означать «housing that someone is living in» или «the country or state or city where someone lives» [1].

Одной из самых открытых проблем при обработке текстов естественного языка является неоднозначность (многозначность) его единиц, что сказывается на всех уровнях и выражается в явлениях полисемии, омонимии и синонимии. Говоря о неоднозначности, можно отметить такие ее виды, как: лексическая, синтаксическая или структурная (например, проблема присоединения - attachment ambiguity), семантическая (когда одно и то же предложение можно по-разному понимать в разных контекстах, хотя лексическая или структурная многозначность отсутствует), прагматическая неоднозначность (когда одно предложение можно по-разному понимать в контексте, в котором оно существует) [2].

Современные системы решения лексической многозначности имеют точность в диапазоне 60-70% и, чаще всего, представлены как самостоятельные методы. Решение проблемы снятия неоднозначности требует интеграции нескольких источников информации и методов. Несмотря на все перечисленные трудности, технология обработки естественного языка в большинстве случаев способна достаточно успешно справиться со своими задачами, поэтому очень полезна во многих отраслях.

Примерно в половине случаев имеет место любая форма омонимии, и набор морфологических признаков оказывается недостаточным для ее решения. Уменьшить неоднозначность можно с помощью синтаксического и семантического анализа с использованием статистических методов, которые и позволяют отбросить крайне маловероятны варианты. Естественный язык хоть и является по своей природе символическим, обработать его с помощью символических, основанных на логике, правил и объективных моделей достаточно сложно. Естественный язык является крайне неоднозначной и переменчивой, поэтому для ее обработки необходимо применять статистические алгоритмы, поэтому доминантными подходами современной обработки языка являются подходы, основанные на статистическом машинном обучении (statistical machine learning) [6].

Машинное обучение (machine learning) исследует изучение и построение алгоритмов, которые могут продуцироваться на основе данных, и выполнять предсказательный анализ на них [7]. Такие алгоритмы действуют путем

построения модели из образцового тренировочного набора входных наблюдений, чтобы создавать управляемые данными прогнозы или принимать решения, выраженные как выходы (результаты) [5], вместо того, чтобы строго придерживаться статических программных инструкций.

Большинство определений машинного обучения отождествляют с наукой о задействовании компьютеров к обучению выполнению действий - подобно как это делают люди для улучшения их обучения в течение долгого времени в самостоятельном режиме, наполнения данными и информацией в виде наблюдений и реального взаимодействия с окружающими. В современных условиях машинное обучение, в основном, развивается в направлении глубокого обучения (deep learning) - это такой тип обучения, в котором модель учится решать задачи классификации непосредственно с изображениями, текстом или звуком. Глубокое обучение, как правило, осуществляется с помощью архитектуры глубокой искусственной нейронной сети.

Основной характеристикой, определяющей преимущество глубокого обучения, является точность. Передовой инструментарий и новейшие методы резко улучшили алгоритмы глубокого обучения. Эти технологии достигли точки, где они могут превзойти людей в классификации изображений, выигрывать против лучших игроков соответствующей отрасли в мире, включать контролируемого голосом помощника от Microsoft Cortana, Amazon Echo или Google Home и пр. Ниже перечислены три технологические возможности, которые позволяют достичь необходимой степени точности глубокого обучения [5]:

1. Легкий доступ к массивам наборов помеченных данных, таких как ImageNet, PASCAL VoC, доступных и удобных для обучения на многих различных типах объектов.
2. Увеличение вычислительных мощностей - графические процессоры (GPU) высокой производительности ускоряют подготовку огромного количества данных, необходимых для глубокого обучения, что позволяет достичь сокращения учебного времени от недель до часов.
3. Предварительно тренированные (Pretrained) модели, построенные экспертами, такие как AlexNet, могут перетренироваться для выполнения новых задач по распознаванию естественного языка (и других даных) и используют технологию, названную «передача обучения». Хотя AlexNet учился на 1,3 млн. образах с высоким разрешением для распознания 1000 различных объектов, но точную передачу обучения достиг с гораздо меньшим набором данных.

Алгоритмы машинного обучения является убедительно привлекательными для внедрения в различные сферы социально-экономического хозяйствования. Насчитывается большое количество проектов на основе технологий машинного обучения. Как отмечает Д. Фаджела, в условиях ведения бизнеса перед стартапами пока возникает вопрос - начинать бизнес-проект с начальным привлечением технологии машинного обучения или начинать его использовать на более поздней стадии развития бизнеса [9]. Следует при этом иметь в виду, что рентабельность инвестиций в машинное обучение требует кропотливой работы по настройке и корректировке.

Доктор университета Оклахомы Данко Николс в своей статье [8] отмечает, что наиболее распространенными ошибками, которые делают предприятия при использовании машинного обучения является мнение о том, что решения систем и принципов машинного обучения для конкретной сферы - одноразовый процесс: менеджеры присылают данные специалистам по данным (data scientists), которые готовят готовые модели. На самом деле поиск хорошего решения - это итерационный процесс, который включает в себя исследования, пробы и ошибки, экспериментирование, консультирование бизнес-специалистов и т.д. Утверждение о том, что машинное обучение никогда не сможет стать товаром, и его успех сильно зависит от знаний, навыков и самоотверженности тех людей, которые это делают [8] - выглядит достаточно убедительным.

Вывод. Под обработкой естественных языков понимают создание систем с признаками искусственного интеллекта, которые определенным образом обрабатывают речевую информацию с целью выполнения определенных задач. К таким задачам относятся: чат-боты или формирование ответов на вопросы пользователя; определение характера эмоциональной окраски высказываний; машинный перевод с одного языка на другой; распознавание языков; проверка правописания; определение частей речи в предложении и их аннотирование; рерайт текстовой информации для создания веб-контента. Машинное обучение в рамках обработки естественных языков представляет собой актуальную сферу научного знания, которая интенсивно развивается и имеет очень большие перспективы. В наиболее узком смысле под машинным обучением понимают класс методов искусственного интеллекта, характерной чертой которых является не прямое решение поставленной задачи, а применение для этого специально обученной математической модели. Такая модель учится за счет решения большого количества задач в нужной области. Смысл применения машинного обучения для обработки естественных языков заключается в том, что глубинные нейронные сети выполняют работу, на

осуществление которой в течение приемлемого промежутка времени нужно было бы применять десятки или даже сотни команд профессиональных лингвистов. Традиционные нейронные сети не имеют возможности принимать текущие решения на основе своих предыдущих суждений. Большое количество задач, решаемых при машинной обработке естественных языков, требует поэтапного анализа данных с учетом предыдущих результатов. Нейронная сеть должна «читать» предложение слово за словом, «осмысливая» его значение исходя из контекста.

Одним из самых перспективных современных технологий машинного обучения является применение глубинных нейронных сетей, в основе которых лежит применение глубокого обучения. Глубокое обучение - это набор алгоритмов машинного обучения, которые позволяют создавать модели с высоким уровнем абстракции в исходных данных, используя архитектуры нейронных сетей, содержащих нелинейные преобразования сигнала. Характерной особенностью алгоритмов глубоко обучения является прохождение входящей информации через гораздо большее количество слоев, чем при традиционном (поверхностном) обучении. Для нейронных сетей с рекуррентной архитектурой путь, которым проходит информационный сигнал от входа к выходу, является теоретически неограниченным (практически он может ограничиваться возможностями примененного программного обеспечения). Благодаря использованию глубинных нейронных сетей решаются задачи компьютерного видения, обработки естественных языков, распознавание музыки, прогнозирования развития событий, интеллектуальной фильтрации данных, построения чат-ботов и др. Большое количество привычных и удобных сервисов компании Google, которые популярные сейчас во всем мире и во всех сферах применения, было бы совершенно невозможным без применения глубинных нейронных сетей. Следовательно, обработка естественных языков посредством применения машинного (глубокого) обучения является перспективным направлением в теории и практики наук современного развивающегося информационного общества.

Литература

1. Баранов Я.В., Радченко И.А., Миронов А.Ю. Использование средств обработки естественного языка для улучшения произношения на иностранном языке // Информатизация образования и науки. 2018. № 3 (39). С. 98-105.
2. Боярский К.К. Введение в компьютерную лингвистику. Учебное пособие. СПб: НИУ ИТМО, 2013. 72 с.

3. Стельмах М.А., Миснянкин В.Г., Кунац А.Ю., Костина А.В. Использование промежуточных языков представления для упрощения процесса перевода естественного языка в запросы к базе данных // Наука настоящего и будущего. 2017. Т. 1. С. 114-116.
4. Юргель В.Ю. Сложности моделирования естественного языка // Вестник науки и образования. 2019. № 23-1 (77). С. 12-14.
5. Dou Z., Wang X., Shi Sh., Tu Zh. Exploiting deep representations for natural language processing // Neurocomputing. 2020. Vol. 38 (621). P. 1-7.
6. Giménez M., Palanca J., Botti V. Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis // Neurocomputing. 2020. Vol. 78. P. 315-323.
7. Hamilton L.M., Lahne J. Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development // Food Quality and Preference. 2020. Vol. 83.
8. Nikolic D. The Human Side of AI // School of Finance and Management (DBIS). Frankfurt: Goethe-University, 2017.
9. Faggella D. What is Machine Learning? // Emerj. URL: https://emerj.com/ai-glossary-terms/what-is-machine-learning/ (дата обращения: 03.05.2020).

Literature

1. Baranov Ya. V., Radchenko I. A., Mironov A. Yu. Use of natural language processing tools to improve pronunciation in a foreign language // Informatization of education and science. 2018. no. 3 (39). Pp. 98-105.
2. Boyarsky K. K. Introduction to computer linguistics. Textbook. Saint Petersburg: ITMO research INSTITUTE, 2013. 72 p.
3. Stelmakh M. A., Misnyankin V. G., Kunats A. Yu., Kostina A.V. Using intermediate representation languages to simplify the process of translating natural language into queries to the database / / Science of the present and future. 2017. Vol. 1. Pp. 114-116.
4. Yurgel V. Yu. Complexity of natural language modeling // Bulletin of science and education. 2019. no. 23-1 (77). P. 12-14.
5. Dou Z., Wang X., Shi Sh., Tu Zh. Exploiting deep representations for natural language processing / / Neurocomputing. 2020. Vol. 38 (621). P. 1-7.
6. Giménez M., Palanca J., Botti V. Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis // Neurocomputing. 2020. Vol. 78. P. 315-323.
7. Hamilton L.M., Lahne J. Fast and automated sensory analysis: Using natural

language processing for descriptive lexicon development // Food Quality and Preference. 2020. Vol. 83.
8. Nikolic D. The Human Side of AI // School of Finance and Management (DBIS). Frankfurt: Goethe-University, 2017.
9. Faggella D. What is Machine Learning? // Emerj. URL: https://emerj.com/ai-glossary-terms/what-is-machine-learning/ (accessed: 03.05.2020).