УДК 004.85

Гарколь Наталья Станиславовна
кандидат педагогических наук, доцент
доцент кафедры педагогики
Английский государственный
педагогический университет
г. Барнаул, Россия

Natalya Garkol
PhD in Pedagogy, Associate professor
Associate professor
of the department of pedagogy
Barnaul, Russia
n_garkol@mail.ru

ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЗАДАЧАХ ЛИНГВИСТИКИ

В статье рассматриваются технологии искусственного интеллекта в лингвистических задачах, технологии машинного перевода текста с одного языка на другой.

К л ю ч е в ы е с л о в а: *технологии искусственного интеллекта; технологии машинного перевода.*

ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN LINGUISTIC TASK

The article discusses artificial intelligence technologies in linguistic tasks, technologies for machine translation of text from one language to another.

K e y w o r d s: *artificial intelligence technologies; machine translation technologies.*

В современном мире технологии искусственного интеллекта (ИИ) с каждым днем играют все более важную роль в решении различных лингвистических задач, таких как обработка естественного языка (NLP), машинное обучение (ML) и глубокое обучение (DL).

На первый план выходят такие компьютерные приложения, которые отвечают за обработку естественного языка, производят извлечение структурированных данных из неструктурированного текста, например, имен, дат и событий; технологии машинного перевода текста с одного языка на другой; генерация естественного языка, когда текст звучит так, как написанный человеком; технологии распознавания именованных сущностей, осуществляющих идентификацию и классификацию именованных сущностей в тексте, таких как имена людей, организаций и мест.

Конечно, использование таких технологий помогает нам в обычной жизни, как в изучении языка, так и в путешествиях в зарубежные страны для упрощения информационно-коммуникативного общения.

Тем не менее мы видим частые ошибки при переводе с китайского языка на русский язык. Для повышения качества перевода современный переводчик может активно пользоваться нейронными сетями, улучшая качество перевода, особенно когда это касается технической терминологии. Здесь цифровые технологии играют важную роль в сфере обучения переводу.

Рассмотрим этапы обработки естественного языка для дальнейшего алгоритма классификация текста.

Для достижения вышеуказанной цели главным является понимание закономерностей в тексте. На первом этапе необходимо научиться выполнять процедуру «токенизация» – разбиение текста на более мелкие части – токены. К токенам относятся и слова, и знаки пунктуации. Токены полезны для нахождения таких паттернов, а также рассматриваются как базовый шаг для дальнейшего анализа. После токенизации обычно создается словарь, в который заносятся уникальные лексемы, встретившиеся в корпусе или тексте.

Но проблема часто заключается в том, что русский язык – это язык с богатой морфологией, имеющий развитые системы склонений и спряжений слов. При работе с текстами на этих языках сложность возникает при составлении словаря, когда нужно найти и объединить все словоформы одной и той же лексемы. Например, *книга – книгу – книгой* – это не уникальные лексемы, а одно и то же слово в разных падежных формах. Здесь применяют процедуру, которую называют «стемминг» (от английского *stem* ‘стебель’), когда у слов просто «отрезают» окончания.

Китайский язык – это язык с продуктивным сложением основ-графем, например: 降尘器 （jiàng chénqì）

降 падать

尘 пыль

器 аппарат

压力陶瓷 (yālì táocí)

压力 давление

陶瓷 фарфор

Современные разработчики искусственного интеллекта в области автоматизации различных лингвистических задач все еще не могут придумать универсальное определение понятию «слово». Мы привыкли к языкам европейского типа, где «слово» – это набор букв между пробелами и знаками препинания. По таким разделителям компьютер можно легко обучить находить слово. Но в китайском языке между словами вообще нет пробелов. Поэтому создание универсального «токенизатора» является сложной задачей.

Но современные лингвисты уже должны уметь на практике пользоваться нейронными сетями для создания таких «токенизаторов».

Например, возьмем исходный текст

该机采用的是柱塞式全自动润滑加油系统，即使在很低的缝速下，仍有很好的供吸油性能，通常进油量除旋梭油量可调外，其余油量均不可调。旋梭的油量，可以调节机油流量调校螺钉来加以控制。首先扳松调整螺丝螺母，当顺时针转动调校螺钉时，旋梭油量增加，反之，则旋梭油量减小，调节好后，把螺母扳紧。

Определим по тексту наборы слов, которые будем использовать для распознавания предложений, где речь идет о механизмах, устройствах, или системных устройствах, т.е. создадим соответствующие словари.

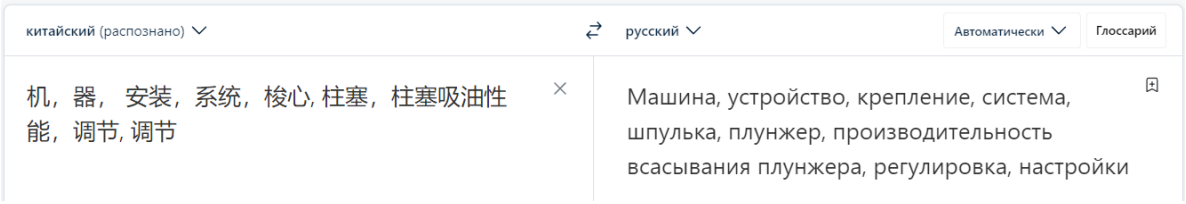
{机，器， 安装，系统，梭心，柱塞，柱塞吸油性能，调节，调节}

Механизм, устройство, установка, система, плунжер, поршень, маслоёмкость, настройка, регулировка}

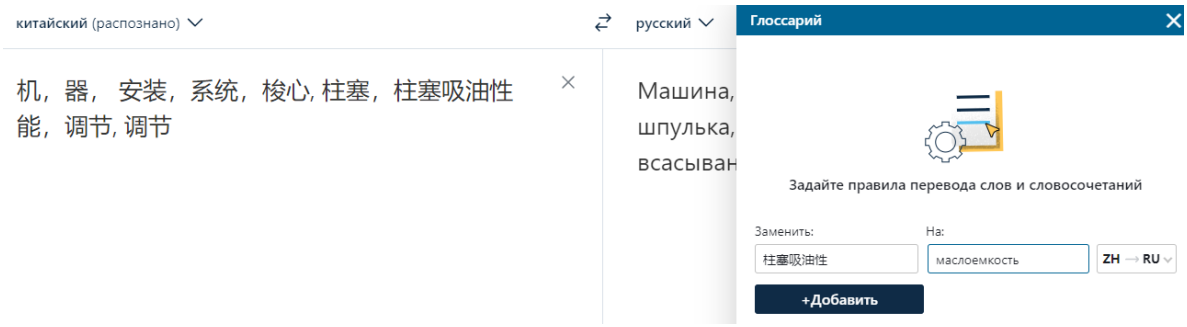
Если «слово» имеет несколько значений, указываем его соответствующее количество раз со всеми возможными техническими терминами.

Таким образом, мы заранее определяем ограниченное числовое пространство решений, характерные признаки данных, помогающие предсказать цель.

Приведем пример работы начального перевода заданного тексту набора слов. Нейронная сеть DeepL выдала следующий результат.

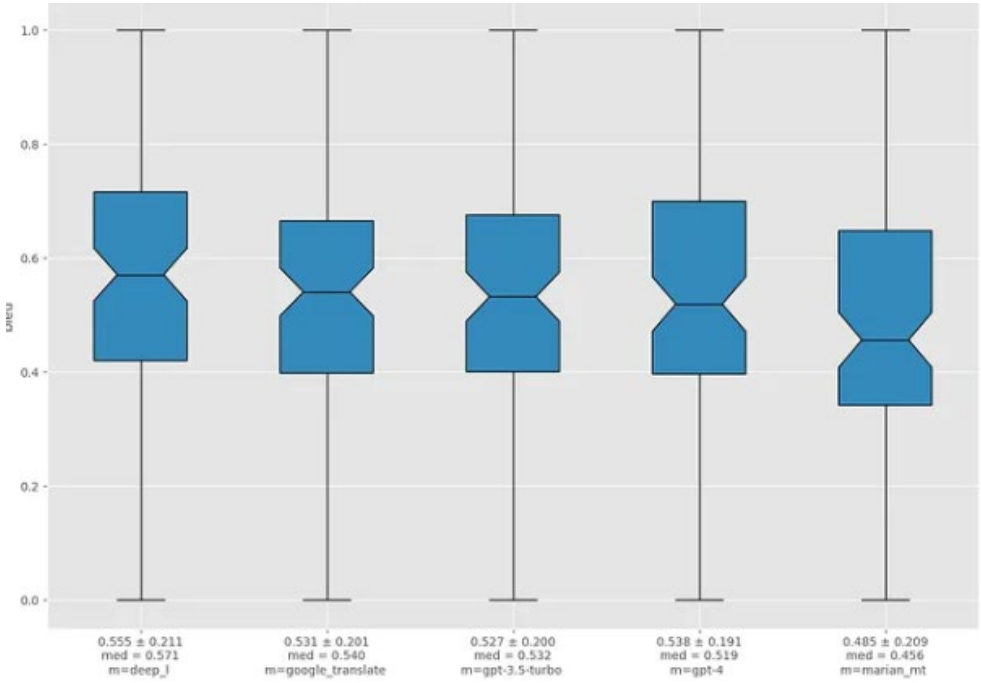


Далее обучаем сеть на сложные термины технического языка.



Обученная система автоматически выдает требуемый результат.

Далее проведем визуализацию данных распределения BLEU для языковой пары «русский–китайский» в виде «ящиков», где утолщенная линия посередине означает медиану, границы ящика показывают 25-й и 75-й перцентили. Треугольный вырез в центре называется “notch” и показывает доверительный интервал для медианы.



Как видим, даже в самых худших переводах дообученные нейронные сети корректно передают смысл.

Они могут использоваться для создания интерактивных учебных платформ, развития приложений и программного обеспечения для автоматизированного перевода, обучения при помощи онлайн-курсов и виртуальных классов, а также для обратной связи и улучшения навыков перевода с помощью специализированных программ. Такие технологии также позволяют студентам получать доступ к онлайн-ресурсам, включая электронные словари, корпуса текстов и другие инструменты, улучшающие навыки перевода.

Классификация текста – это задача машинного обучения, которая заключается в назначении категорий или классов текстовым данным.

Технологии ИИ трансформируют область лингвистики, предоставляя мощные инструменты для решения сложных лингвистических задач. По мере развития этих технологий мы можем ожидать дальнейших инноваций и прорывов в этой области.

ЛИТЕРАТУРА

1. Искусственный интеллект в лингвистике: основные задачи и методы для оптимизации сайтов [Электронный ресурс] // Научные Статьи.Ру : портал для студентов и аспирантов. URL: https://nauchniestati.ru/spravka/ii-v-lingvistike/ (дата обращения: 13.03.2024).

2. Ермоленко О., Горев И. И. Математические методы в лингвистике: применение модели нейронных сетей в системе машинного перевода // Переводческий Дискурс: междисциплинарный подход : материалы 3- международной научно-практической конференции. 2019 г. (дата обращения: 13.03.2024).