СЕКЦИЯ 5. ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА: ОПЫТ И ПЕРСПЕКТИВЫ

УДК 371.3

Авраменко Анна Петровна
кандидат педагогических наук, доцент
МГУ имени М.В. Ломоносова
Москва, Россия

Anna Avramenko
PhD in Pedogogy, Associate Professor
Lomonosov Moscow State University
Moscow, Russia

ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ НАУЧНЫХ СТАТЕЙ

Сегодня технологии искусственного интеллекта (ИИ) позволяют эффективно и в автоматическом режиме обрабатывать большие массивы текстов. Таким образом функционируют национальные корпусы, а по их подобию исследователи проводят лингвистический анализ подкорпусов текстов по своим научным интересам. В рамках дисциплины по выбору «Основы обработки естественного языка технологиями ИИ» для бакалавров (при поддержке Фонда «Интеллект») и курса «Компьютерные технологии в лингвистических исследованиях» на факультете иностранных языков и регионоведения (ФИЯР) МГУ имени М. В. Ломоносова студентам для защиты итогового проекта предлагается составить и проанализировать подкорпус актуальных научных статей по тематике своего интереса для выявления основных тенденций по частотным языковым единицам. В данной статье предлагается описание технологий, лежащих в основе такого подхода.

К л ю ч е в ы е с л о в а: *искусственный интеллект; обработка естественного языка; лингвистический анализ; корпусная лингвистика; современные исследования.*

ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR LINGUISTIC ANALYSIS OF TEXTS OF SCIENTIFIC ARTICLES

Today, artificial intelligence (AI) technologies allow processing big text data efficiently and automatically. National corpora function in this way, and in their likeness, researchers conduct linguistic analysis of sub-corpora of texts according to their scientific interests. Within the framework of the elective discipline “Basics of Natural Language Processing with AI Technologies” for bachelors (supported by the Intellect Foundation) and the course “Computer Technologies in Linguistic Research” at the Faculty of Foreign Languages and Area Studies (FFLAS) of Lomonosov Moscow State University, students are invited to compile and analyze their sub-corpus of up-to-date scientific articles on the topic of their interest in order to identify the main trends in the frequency of linguistic units. This paper offers a description of the technologies underlying this approach.

K e y w o r d s: *artificial intelligence; natural language processing; linguistic analysis; corpus linguistics; modern research.*

Существуют различия в подходах к созданию и использованию корпусов в России и за рубежом. В первой половине XX века американскими структуралистами были заложены основы корпусной лингвистики как

эмпирической методологии. Основным критиком такого подхода является Ноам Хомский [1]. Так, критика использования больших данных в лингвистическом исследовании связана прежде всего с тем, что количественные показатели не всегда отражают качественные данные. Индикатор частотности использования одного слова не отражает, каким словом или выражением пользовались до этого. Развитие *корпусной лингвистики* как отдельного направления относится к концу XX века. Именно в этот период лингвисты начали объединять полученные данные в рамках размеченных массивов лингвистических данных в электронных корпусах.

В. П. Захаров определяет *лингвистический корпус* (corpus, множественное число – corpora) как большой, представленный в электронном виде, структурированный и размеченный, представительный массив языковых данных, предназначенных для решения определенных лингвистических задач [2]. Благодаря применению корпусов можно, например, сделать выводы о лексическом фонде устойчивых словосочетаний и об особенностях их использования. С момента появления электронные лингвистические корпусы находили свое применение в двух направлениях: с одной стороны, получение примеров использования определенной языковой единицы и представления о частотности ее употребления; с другой – решение задач компьютерной лингвистики, в том числе машинное обучение.

Навигация в электронном лингвистическом корпусе происходит с помощью корпусного менеджера, посредством которого осуществляется обработка статистической информации с ее последующим предоставлением пользователю. Такие программы называют *конкордансами*, поскольку они позволяют выстроить список всех словоупотреблений в контексте со ссылками на источник (concordance). Современный *корпусный менеджер* должен выполнять следующие задачи: строить конкорданс, искать контексты по словам и словосочетаниям (n-граммам, биграммам, триграммам и т. д.), сортировать списки по нескольким критериям, анализировать словоформы, давать статистическую информацию и метаданные о словоупотреблении (о том, что включают метаданные, речь пойдет ниже). Двумя наиболее популярными системами для обработки авторских корпусов являются AntCont и SketchEngine [3].

В 2004 году в России был запущен *Национальный корпус русского языка (НКРЯ)*. Согласно данным официального сайта корпуса на конец 2023 года он представляет из себя коллекцию текстов на русском языке общим объемом более 2 млрд слов. Ресурс ruscorpora с инструментами поиска разрабатывается компанией Яндекс. В 2023 году НКРЯ был значительно обновлен и расширен, в том числе сегодня доступны: панхронический корпус (включающий в себя тексты трех исторических корпусов: древнерусского, старорусского и корпуса берестяных грамот); корпус «Русская классика» (в том числе с черновиками и редакционными версиями произведений);

корпус «От 2 до 15» (с популярными сегодня среди детей и подростков произведениями); актуальный корпус «социальные сети», для которого использована модель RuRoBERTa; а также несколько десятков параллельных корпусов (например, параллельный русско-китайский подкорпус НКРЯ развивается с 2016 года и насчитывает 4,5 млн слов и более тысячи документов разных жанров и стилей).

Примером международного корпусного проекта является диахронический корпус *Google Ngram Viewer* на 9 языках. Библиотека размеченных текстов русского языка Google Books составляет более 0,5 млн документов. Данный инструмент имеет ряд уникальных функций в пользовательском интерфейсе: обработка биграмм (n-грамм из двух слов, или словосочетаний/коллокаций), учет позиции слова в предложении и гибкая работа с графиками.

Выделим некоторые *задачи,* которые большие языковые модели как средства обработки естественного языка (Natural Language Processing, NLP) могут решать в рамках лингвистических исследований:

- классификация, или присвоение метки класса объектам (например, как с определением изображений, для распознавания речи используется классификация звуков по тем или иным категориям);
- кластеризация, или распределение на группы (например, выделение тем в корпусе текстов);
- ранжирование, или сортировка по признакам (например, для определения релевантности поисковой выдачи).

Наиболее сбалансированным решением для выполнения вышеперечисленных задач представляется использование *предварительно обученной на большом корпусе данных модели с последующей дополнительной настройкой ее с помощью доступных данных.*

Дообучение большой языковой модели под конкретные задачи приложения происходит следующим образом. После выбора архитектуры нейроны на базе больших данных дают предсказание и вычисляют ошибку как разницу между предсказаниями и верными результатами. В 2020-е оптимальным решением становится именно дообучение предобученных трансформеров так называемыми методами Fine tuning и Transfer learning, поскольку данный процесс не требует дополнительных объемов размеченных больших данных (о разметке текстовых данных речь пойдет в разделе, посвященном корпусной лингвистике). Для работы с предобученными трансформерами используется ряд общепринятых *технологий:*

- *инфраструктурные*

1) *язык программирования* Python (появился в начале 1990-х, получил известность с 2019 года с развитием нейронных сетей; один из самых простых языков программирования и самый подходящий для машинного обучения); для более сложных задач обработки данных может применяться язык R;

2) среда обработки, иными словами, *блокнот* Jupyter (Integrated Development Environment, IDE), где можно писать код; это может быть программное обеспечение (ПО) для компьютера, как PyCharm, или онлайн-аналоги, как Google Colab и Yandex.DataSphere для исследователей;

• *программные* (приведенные ниже примеры программного обеспечения представляют из себя открытые библиотеки на ресурсе GitHub, а обращение к ним происходит через копирование с GitHub и добавление в свой блокнот их кода API, Application Programming Interface):

3) в качестве *базы данных* (БД) для задач обработки естественного языка выступает лингвистический корпус или подкорпус, состоящий из документов; где документ – это совокупность токенов, которые принадлежат одной смысловой единице (в качестве документа может выступать предложение, комментарий или пост пользователя); например, наиболее известные датасеты на русском языке – это deepavlov;

4) *система управления базой данных* (СУБД), например, SQLight; используется для реляционных баз данных (то есть таблиц) на основе языка запросов SQL (Structure Query Language);

5) открытые *библиотеки* Pandas (базовая библиотека по подготовке данных) + NumPy (основная библиотека для обучения модели, на ней разработан пакет ScikitLearn) – обычно используются в паре для машинного обучения в целом; а NLTK (Natural Language Toolkit) – это наиболее популярная международная библиотека для задач обработки текста (есть модули для русского языка + можно применять ПО от Яндекс для обработки естественного языка MyStem);

6) *фреймворки* TensorFlow или PyTorch делят примерно пополам 90 % рынка приложений на базе машинного обучения (еще иногда используется Keras).

Нами был разработан алгоритм обработки подкорпуса научных статей для выявления тенденций в определенной области на основе обработки частотных языковых единиц открытыми библиотеками. С помощью данного алгоритма в рамках вариативной части программ бакалавриата и магистратуры мы предлагаем студентам собирать данные для теоретической части их выпускных квалификационных работ. Перспективным представляется изучение потенциала технологий обработки текстов в технических дисциплинах.

ЛИТЕРАТУРА

1. Chomsky N. On nature and language. Cambridge, Ma : MIT Press, 2002. 206 p.

2. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. Изд. 3. СПб., 2020. 235 с.

3. Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, И. В. Митрениной, Т. М. Ландо. М., 2016. 320 с.