НЕЙРОННЫЕ СЕТИ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ЛИНГВООБРАЗОВАНИИ, НАУКЕ И ПРОИЗВОДСТВЕ

М.Б. Чиков
к.ф.н., доцент старший научный сотрудник
научно-образовательного центра
дизайна и проектирования инновационной среды
дополнительного образования
Нижегородского государственного лингвистического университета
им. Н.А. Добролюбова
Нижний Новгород, Росссия
chikov@lunn.ru

Д.Н. Зарипова
и.о. доцента Ташкентского государственного университета
узбекского языка и литературы им. Алишера Навои
Ташкент, Узбекистан
dilyaran78@gmail.com

Н.Д. Чикова
старший преподаватель кафедры международного
менеджмента, экономики и информационной безопасности
Нижегородского государственного лингвистического университета
им. Н.А. Добролюбова
Нижний Новгород, Росссия
nanachikova@yandex.ru

Л.С. Павлова
к.п.н., доцент кафедры методики преподавания
иностранных языков, педагогики и психологии
Нижегородского государственного лингвистического университета
им. Н.А. Добролюбова
Нижний Новгород, Росссия
lszhukova@lunn.ru

DOI: https://doi.org/10.5281/zenodo.18229048

Аннотация. *Научно-образовательные и производственные лингвистические процессы в настоящее время невозможны без искусственного интеллекта, в частности, представленного нейронными сетями. Зарубежные нейронные сети успешно решают вопросы обработки даже экзотических языков, что говорит о необходимости создания такой сети в России (в том числе в сотрудничестве с Республикой Узбекистан). Обработка материала редких языков России, в свою очередь, пока не доступна никаким сетям, что требует от российских исследователей больших усилий в создании такого инструмента. Программа «Фразеоэкстрактор», созданная учеными НГЛУ на базе традиционных средств программирования, для решения указанных задач должна быть реализована в формате нейронной сети.*

Ключевые слова: *искусственный интеллект, нейронная сеть, модель, трансформер, набор данных, машинное обучение, узбекский язык, редкие языки России, автоматизация перевода.*

Нейронные сети, искусственный интеллект не могут не затронуть современные процессы в изучении и сравнении языков, обучении переводу и переводческой деятельности.

В Нижегородском государственном лингвистическом университете им. Н.А. Добролюбова на протяжении нескольких последних лет ведется работа над современным инструментом извлечения языкового материала из текстов.

Известны традиционные средства обработки естественного языка, например, библиотека *Spacy* языка программирования *Python*, которая содержит материал нескольких десятков европейских языков и позволяет обрабатывать их материал достаточно эффективно. Классическая обработка языков включает частеречную разметку, извлечение групп из двух слов, связанных определенными отношениями (глагол и прямое дополнение, существительное и определение и т.п.), извлечение именованных сущностей (*named entities*; проще говоря, имен собственных с их классификацией – географические названия, имена лиц, названия организаций). Наиболее актуальным нам представляется извлечение словосочетаний различных типов из текстов различных жанров и тематик. Это дает возможность получать самый актуальный языковой материал, на котором непосредственно основывается говорение и перевод. Так как в традиционные средства заложена, как сказано выше, только модель извлечения словосочетаний из двух слов по принципу семантико-грамматического отношения между ними, а типов словосочетаний гораздо больше и слов в словосочетаниях, как правило, не два, а три – четыре или в ряде случаев даже больше, мы работали средствами библиотеки *Spacy* над возможностью извлечения всех потенциальных словосочетаний объемом до шести слов, используя вероятный порядок следования частей речи друг за другом в разных языках. Нельзя не отметить, что, например, в немецком языке этот принцип сталкивается с серьезным препятствием, заключающимся в разном положении глагола и его отделяемой приставки относительно других слов в предложении. Таким образом, выстраиваемая нами модель с глагольными конструкциями немецкого языка не справляется, а встроенная модель *Spacy*, извлекающая глагол и прямое дополнение (независимо от позиции), как мы уже говорили, как правило, не дает полных словосочетаний, а также не захватывает отделенную приставку.

Ввиду отсутствия модели обработки узбекского языка нами была создана пустая модель библиотеки *Spacy*, которая была обучена на минимальном наборе данных и позволила сделать первые извлечения словосочетаний из простых текстов. Следует отметить, что модель справлялась не только с обработкой слов, присутствующих в наборе данных, но и интуитивно верно

определяла некоторые существительные (особенно по суффиксу множественного числа *-lar*), некоторые глаголы (вероятно, также по каким-либо характерным признакам), интернационализмы.

Об истории разработки нашего фразеоэкстрактора см. статьи Е.А. Кабановой [1, 2, 3, 4], М.Б. Чикова [5], Н.Д. Чиковой, Д.Н. Зариповой [6].

Передовые нейронные сети, например, *DeepSeek*, хорошо обрабатывают узбекский язык (в отличие от *YandexGPT*). Таким образом, наша следующая задача – создание нейронной сети, трансформера, обученного в первую очередь для лингвистического анализа. Трансформеры используют так называемый механизм внимания, т.е. одновременного восприятия информации на разных входах, что и делает их эффективнее предыдущих нейронных сетей. Так, например, трансформеры архитектуры *BERT* (*bidirectional encoder representations transformer*, двунаправленный кодировщик-трансформер) рассматривают единицы в тексте (слова) одновременно и с левым, и с правым контекстом и таким образом достигают точности в определении характеристик каждой единицы. Над такой моделью в настоящее время и ведется работа. Если в случае с европейскими языками, для которых уже существуют большие размеченные наборы данных, обучение этой модели не будет представлять трудностей, то для узбекского языка такого материала не существует. Отметим, что у создателей таких сетей, как *DeepSeek*, которые, как было сказано, отлично справляются с обработкой узбекского языка, необходимые наборы данных есть, но их нет в открытом доступе. Поэтому для обработки узбекского языка – а этот язык в настоящее время представляет для нас первостепенный интерес – набор данных необходимо создавать с нуля. В отличие от эксперимента со *Spacy* и минимальным набором данных, этот процесс в настоящее время несколько систематизирован. Наборы данных узбекского языка есть, но они не размечены, т.е. они содержат тексты, но слова не имеют лингвистических пометок (часть речи, грамматические признаки и т.п.). Модели для разметки узбекского материала также нет. Ввиду этих обстоятельств мы взяли первоначальный набор данных из тысячи текстов (114 000 слов, в том числе повторяющихся; количество без повторов не подсчитывалось) и разметили этот набор частеречными признаками при помощи модели турецкого языка библиотеки *Stanza*. Разумеется, турецкая модель почти безошибочно определяет существительные, особенно если они имеют суффикс *-lar*, а в глаголах, которые в узбекском языке имеют достаточно необычную структуру, часто ошибается. Таким образом, сейчас идет работа над ручной правкой предварительно размеченного датасета; разумеется, такая автоматизированная работа продвигается быстрее, чем полностью ручное создание. При достижении

количества правильно размеченных слов 3000 на этом наборе данных будет обучена модель *BERT*, которую затем можно будет использовать для анализа текста (определение части речи слов в тексте) и дообучать на последующих данных.

Кроме обработки узбекского языка, есть задачи, которые пока не могут решить и передовые нейронные сети и которые крайне актуальны для России: это компьютерная обработка малых языков России, особенно находящихся на грани исчезновения. В настоящее время принимаются активные меры для изучения и сохранения нанайского языка и нанайской культуры (тунгусо-маньчжурский язык; центр нанайского языка существует в Приамурском государственном университете), планируется начать изучение языков Красноярского края во взаимодействии с его университетами, в первую очередь Сибирским федеральным университетом: эвенкийский (тунгусо-маньчжурский), представленный как в Приамурье, так и в Красноярском крае, кетский (последний енисейский язык, локализованный в Красноярском крае и находящийся под угрозой исчезновения) и другие.

Для данных языков необходимо внедрить как стандартные виды обработки – частеречную разметку, определение грамматических признаков, извлечение именованных сущностей, извлечение словосочетаний, определение тональности текста – так и создать базы данных для сохранения и изучения материала, системы автоматизации перевода, лингвотренажеры для освоения языкового материала редких языковых семей и т.д. Понятно, что наборы данных для этих языков будут создаваться полностью с нуля, так как, даже в отличие от узбекского языка, для этих языков их не существует в принципе и нет модели для разметки, которая хотя бы приблизительно была похожа на необходимую, как это было в случае с узбекским языком и турецкой моделью.

Почему из всех языковых единиц мы в первую очередь уделяем внимание словосочетаниям? Это связано с тем, что словосочетания представляют собой готовые предикативные группы, из которых быстро, легко и правильно составляется живая речь. На актуальных словосочетаниях должны строиться учебники, курсы, лингвотренажеры, терминологические базы. Последнее особенно важно для переводчиков, которые работают в системах автоматизации перевода. Перед выполнением собственно перевода они готовят двуязычную терминологическую базу; еще раз подчеркнем, что термины должны быть не словами, а словосочетаниями (сравним: *solid* как термин имеет одно значение, а *solid engine* – другое). Извлечение словосочетаний из текста оригинала лучше всего способна произвести нейронная сеть, обученная для решения этой задачи. Она же может дать перевод этих терминов. После

недолгой проверки на соответствие переводчик или терминолог загружает полученную таблицу в систему автоматизации перевода, и все участники проекта имеют возможность ориентироваться на нее (термины вставляются в текст щелчком мыши).

Таким образом, нейронные сети и искусственный интеллект уже стали участниками всех научно-образовательных и производственных лингвистических процессов, но они должны получить еще большее распространение и, главное, качественная нейронная сеть должна быть именно *нашей* для обеспечения независимости от других государств, от нестабильности всех Интернет-систем вообще, и для решения задач, которые пока не решаются ничем.

Литература

1. Кабанова Е.А. Фразеология специального текста (на материале немецкого языка) // В сборнике: VII Международная научно-практическая конференция «Гармонизация межнациональных отношений в условиях глобального общества», XXVI Нижегородская сессия молодых ученых (гуманитарные). Сборник статей и тезисов молодых ученых. Нижний Новгород, 2021. – С. 289-292.
2. Кабанова Е.А. Цифровые инструменты для лингвистического анализа и обработки словосочетаний на базе немецкоязычных медицинских текстов // VIII Международная научно-практическая конференция «Гармонизация межнациональных отношений в условиях глобального общества», XXVII Нижегородская сессия молодых ученых (гуманитарные науки). Нижний Новгород, 2022. – С. 340-344.
3. Кабанова Е.А. Современные инструменты изучения фразеологии // Непрерывное образование 4.0: вызовы, тренды и стратегии подготовки кадров будущего. Сборник научных статей по материалам международной научно-практической конференции. Под редакцией Ю.В. Чичериной, М.Б. Чикова. Нижний Новгород, 2022. – С. 52–56.
4. Кабанова Е.А. К вопросу о классификации словосочетаний (на материале немецкого языка) // Проблемы языка и перевода в трудах молодых ученых. 2022. – № 21. – С. 69–76.
5. Кабанова Е.А., Чиков М.Б. Создание цифрового инструмента сбора и анализа словосочетаний для научно-практических целей // Проблемы языка и перевода в трудах молодых ученых. – 2023. – № 22. – С. 106–114.
6. Чиков М.Б., Зарипова Д.Н., Чикова Н.Д. Подготовка переводчиков и перевод в тройке языков русский / английский / узбекский: языковые аналогии и цифровые инновации // Научный журнал "Современные лингвистические и методико-дидактические исследования". – 2025. – № 1 (65). – С. 91–104.