Подготовка исходных данных для обучения нейросетей
(на примере марийского языка)

А.В. Чемышев,
научный сотрудник МарНИИЯЛИ им. В.М. Васильева

В данный момент только 17 языков в мире неуязвимы: люди, разговаривающие на этих языках, распоряжаются 90% ресурсов Земного шара. Остальные 6000 языков, по другим данным почти 7000 языков можно отнести к так называемым малоресурсным или миноритарным языкам, несмотря на то что в мире около 200 государств. При этом количество языков каждую неделю уменьшается как минимум на один-два языка. По оценкам специалистов, в дальнейшем языки будут умирать с более высокой скоростью: многие предполагают, что скоро в мире останется не более 200 языков. Чем это вызвано? После изобретения книгопечатания Гутенбергом в 15 веке языки начали делиться на письменные и бесписьменные, этот процесс проходил довольно долго, около 500 лет. Даже спустя 400 лет язык мог оформиться в письменный и совершить своеобразный прорыв: на этих языках за последнее столетие публикуются десятки тысяч изданий, так или иначе развивается литературный язык, ведётся преподавание в учебных заведениях и т. д. Но после Карибского кризиса 60-х годов прошлого столетия противоборствующие стороны задумались о повышении катастрофоустойчивости своих пунктов управления. Так появился интернет. Долгое время интернетом пользовалось только военное ведомство США. Когда стало понятно, что ядерной войны не будет, в интернет «пустили» гражданских людей. В дальнейшем туда подключились другие государства: вначале только англоязычные. А потом начались проблемы: изначально обмен данными в тогдашнем интернете предполагал только использование букв английского алфавита, управляющих символов и знаков препинания. Подключение к интернету европейских стран, вызвало необходимость «косметического ремонта» в кодировке, а когда к нему подключился Советский Союз во время своего заката, затем Россию и все страны мира, стало понятно, что нужна единая кодировка, которая будет правильно отображать знаки алфавитов всех языков мира. Так в начале 1990-х годов появился юникод. Но шрифтов, которые поддерживали эту кодировку, были единицы. К тому же производители многих операционных систем бойкотировали юникод. Но книги и периодику, например в России, надо было издавать на марийском, татарском, башкирском, удмуртском, якутском, бурятском и др. языках. По всей России стали использоваться нестандартные, а зачастую самодельные шрифты.

В начале 21 века, когда интернет стал массовым, пользователи, работающие с миноритарными языками, резко ощутили цифровое неравенство. Использование родного языка в глобальной сети для представителей многих языков стало невозможным: некоторые буквы их регулярного алфавита не имели юникод, а значит, использование этого языка в интернете было невозможно. Стало понятно, что эти языки в новых реалиях обречены.

В апреле 2011 года при поддержке Министерства регионального развития Российской Федерации и Совета Европы в Йошкар-Оле была проведена международная конференция: «Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы». Данная конференция собрала на одной площадке специалистов в области использования языков в ай-ти и компьютерных технологиях, были выработаны итоговые документы. Именно тогда были сформулированы минимально необходимые технические инструментарии, обеспечивающие полноценное функционирование языка в едином информационном пространстве, а именно:

1) стандартизованное электронное представление графических систем национальных языков и раскладки клавиатуры;
2) электронные словари;
3) системы проверки правописания;
4) системы автоматического перевода;
5) поисковые системы.

Также было отмечена необходимость создания образовательных систем, обучающих компьютерных программ, электронных учебников, контента на национальных языках, локализации программного обеспечения, и др.

Необходимо отметить, что участники конференции в пункте «создание поисковых систем» имели в ввиду не поиск методом перебора слов и выражений, а поиск запросов, в том числе по смыслу. Следует ради справедливости сказать, что данная идея в полной мере не реализована ни для одного языка в мире. Хотя работы специалистами в области когнитивной лингвистики ведутся: в данный момент с помощью так называемого сентимент-анализа стало возможным определить тональность запроса (предложения): отрицательная, нейтральная и положительная тональность.

После указанной конференции прошло более 10 лет. Время идёт. Эти необходимые 5 пунктов дополнились как минимум тремя:

6) синтезатор речи;

7) распознавание речи:

8) использование искусственного интеллекта для обработки неструктурированных данных.

Использование нейросетей намного облегчило создание автоматических переводчиков, синтезаторов и инструментов распознавания речи и др.

В 2016 году при Республиканском центре марийской культуры был создан отдел инновационного развития марийского языка и культуры. Первой задачей отдела стало включение марийского языка (лугового и горного литературных стандартов) в Яндекс.Переводчик. По просьбе отдела машинного перевода компании «Яндекс» начали готовить исходные данные: формализованную грамматику и параллельный корпус, включающий в себя выравненные русские и марийские предложения. К началу запуска марийского переводчика удалось собрать базу из 250 тыс. выравненных предложений. Но для хорошей работы машинного переводчика необходимо 1 млн. предложений. В 2017 году у «Яндекса» поменялись приоритеты, работа с национальными языками России затормозилась, а марийский язык в Яндекс.Переводчике так и до сегодняшнего дня остался со значком «бета», то есть не доделанным. (Теперь спустя 4 года мы понимаем, что работу надо было непременно продолжать. Но об этом ниже).

С конца 2017 года по 1 января 2021 года отдел инновационного развития марийского языка и культуры Республиканского центра марийской культуры (2 чел.) и МарНИИЯЛИ им. В.М. Васильева (1 чел.) занимались созданием корпуса марийского языка, состоящего минимум из 20 млн. словоупотреблений. Работа подходила к финишу. Во второй половине 2020 года ФНКА Марийцев России на средства субсидии ФАДН инициировала работу по созданию синтезатора марийской речи. Было закуплено необходимое оборудование: сервер с видеокартой для разворачивания нейросети, профессиональная звукозаписывающая техника для студии звукозаписи. Для обучения нейросети можно использовать готовые библиотеки TensorFlow, PyTorch, MXNet. Программа для запуска обучения нейросети состоит из нескольких строчек кода на языке Python. Ну, 2 или максимум 3% от всего объёма работ по созданию синтезатора речи. Зная об этом, тогда, в 2017 году, мы бы не остановили работу по подготовке параллельного корпуса. И к сегодняшнему дню собрали бы все необходимые 1 млн. предложений. Опыт приходит со временем.

Итак, самая сложная работа: подготовка исходных данных. На сайте mari-lab.ru мы информируем общественность о состоянии работ по созданию аудиокорпуса, необходимого для обучения нейросети. В данный момент собрано около 13 часов. Мы планируем собрать к осени этого года 20 часов записи голоса человека, чьим голосом будет говорить наш синтезатор речи. Далее обучаем нейросеть со скрытыми моделями Маркова и получаем наш первый марийский синтезатор речи.

Для распознавания речи также создаётся аудиокорпус. Он должен делаться следующим образом: один и тот же текст озвучивается не в студийных условиях на диктофон. Источники голоса: дети, молодые люди, мужчины, женщины, люди пожилого возраста, в том

числе говорящие с акцентом и с дефектом речи. В аудиофайлах должны присутствовать шумы окружающей среды, транспорта и т. д. Чтобы при обучении нейросеть научилась понимать голос человека, говорящего на марийском языке и переводить его в текст.

Итак наш скромный совет всем языковым активистам России, центрам компетенции и другим заинтересованным лицам, желающим продвигать свои языки в едином информационном пространстве, ай-ти и компьютерных технологиях, а именно желающим создавать машинные переводчики, синтезаторы речи и распознавание речи: даже если у вас нет высококвалифицированных программистов, разбирающихся в нейросетях, даже если у вас нет серверов с хорошими видеокартами, позволяющими разворачивать нейросети, не отчаивайтесь. Это всего лишь 2–3% от всего необходимого объёма работ. Самое главное — это исходные данные. Где взять исходные данные для создания машинных переводчиков? Сканируйте и распознавайте переводную литературу: с 1930-х годов многих классиков русской и мировой литературы регулярно переводили на национальные языки: А.С. Пушкин, Н.А. Некрасов, А.П. Чехов, И.С. Тургенев и др. Для создания паралллельных текстов русские тексты классиков всегда можно найти в интернете. А с 1950-х годов начали активно переводить на русский язык книги на национальных языках России. Например, для марийского языка — это М. Шкетан, Шабдар Осып, С.Г. Чавайн, А. Юзыкайн и др. Эти книги также необходимо сканировать и распознавать. Список переводной литературы обычно фиксируется и хранится в национальных библиотеках регионов. Также во многих национальных республиках переводится официальные тексты, законы, нормативные акты на национальные языки. Для создания базы параллельных текстов можно использовать программы Abbyy Aligner, LF Aligner или писать свои скрипты. После того как будет собрана база из 1 млн. параллельных предложений, этими исходными данными с радостью воспользуется и Яндекс, и Google… К тому моменту у вас могут появиться свои программисты. Или вы можете попросить кого-то из других регионов обучить нейросети и создать свой машинный переводчик.

Работа по подготовке исходных данных для синтезатора речи аналогичная. Создавайте свои студии звукозаписи (аренда коммерческих студий вам влетит в копеечку, и итоге дороже, чем расходы на строительство своей собственной студии). Начинайте писать аудиокниги, создавайте подкасты, озвучивайте электронные учебники. В качестве людей, озвучивающих произведения, выбирайте артистов национальных театров, работников радио и телевидения. Больше по объёму записывайте людей, чьи голосом должен заговорить ваш синтезатор. Желательно иметь и мужской, и женский голос. Роль синтезатора речи в данный момент неоценима: это использование в обучающих программах, образовательных системах. Также нельзя забывать о людях с ограниченными возможностями (имеющих слабое зрение, незрячих). Ведь для них это зачастую единственная возможность узнавать новости на информационных порталах, ознакомиться с произведениями местных авторов.

Распознавание речи могли бы использовать журналисты, пишущие на национальных языках. Записанные многочасовые интервью на диктофон можно будет мгновенно переводить в текст.

Если освоить элементы искусственного интеллекта, то уже можно создавать голосовые помощники на национальных языках, где основным элементом данных систем являются синтезатор речи и распознавание речи. Но это уже в перспективе.

А пока засучите рукава и начинайте собирать исходные данные для нейросетей, как иногда их называют специалисты — «сырьё»: готовьте базу параллельных текстов, записывайте аудиокниги, которые в дальнейшем станут основой аудиокорпуса.

Кстати, базу параллельных текстов можно загружать в качестве «памяти переводов» для облачных систем автоматизации переводов, тем самым облегчая работу переводчиков с русского языка на ваш и с вашего языка на русский и убыстряя их работу в 3–4 раза.