Пресс-релиз

Независимая группа энтузиастов разработала онлайн-переводчик для вепсского языка. Это язык одного из коренных малочисленных народов России — вепсов. Последняя перепись населения насчитала всего 2713 чел., знающих этот язык, и большинство из них живёт на востоке Ленинградской области, на западе Вологодской и на юге Карелии. Благодаря новой разработке люди, интересующиеся вепсским языком и работающие с ним, впервые получили возможность переводить тексты с вепсского языка на русский и с русского языка на вепсский в автоматическом режиме, что может оказать серьёзную поддержку в деле изучения и сохранения этого языка для будущих поколений.

Команда состояла из трёх человек: автора идеи Максима Мигукина, программиста Алексея Куташова и лингвиста Максима Кузнецова. О том, как проходила работа над проектом и какой вклад в это дело внёс каждый из них, расскажут они сами.

Максим Мигукин,
автор идеи

Несколько лет назад я заинтересовался ДНК-генеалогией и своими корнями. Получив результаты ДНК-теста, я узнал, что я на 15% вепс. Дальше были поездки на историческую родину, составление генеалогического древа, изучение всевозможных источников по вепсской культуре и языку, в результате чего я оказался на краткосрочных курсах вепсского языка в Петербурге, которые организовывала «Вепсская община», где я познакомился с преподавателем вепсского языка Максимом Кузнецовым.

После этих курсов я захотел какого-то участия в сохранении вепсской культуры и если идея издавать книги, и альманахи ещё ждёт своего часа, то идея с онлайн-переводчиком всё сильнее манила меня. Я нашёл выходы на ребят из команды Яндекс Translate, безуспешно пытаясь заразить их идеей добавления вепсского к общей базе языков, ещё не понимая, какой пласт работы стоит за этой просьбой. В итоге я нашёл форум Libretranslate, где на мой призыв попробовать сделать онлайн-переводчик для вепсов откликнулся Алексей Куташов.

Всю фактическую, программную работу по созданию переводчика взял на себя Алексей, а всю лингвистическую поддержку — Максим Кузнецов. Я со своей стороны мог лишь искать вепсские тексты в свободном доступе, запрашивать тексты у журналистов и вручную обрабатывать эти тексты так, чтобы они удовлетворяли техническим требованиям для ML.

Да, перед нами ещё много работы для доведения переводчика до ума, но результат уже можно пощупать. И это доказывает, что команда ранее незнакомых между собой людей за несколько месяцев может на голом энтузиазме сделать то, чего не было у интернет-гигантов или серьёзных институтов.

Максим Кузнецов, *лингвист, языковой активист, научный сотрудник лаборатории многофакторного гуманитарного анализа и когнитивной филологии Казанского научного центра РАН*

Я включился в этот проект потому, что как языковой активист всё время занимаюсь разными проектами, связанными с языками народов России, направленными на их популяризацию, так как комбинация «могу», «хочу» и «это будет полезно» лично для меня складывается в мотивацию «это надо делать». А в плане именно переводчика – как на одной из встреч сказал премьер Кыргызстана Жапаров, «языки, которые не попадут в Интернет, ждёт судьба языков, которые не попали в письменность». Если какие-то программные продукты существуют на крупных языках мира и крупных языках России, то почему бы и малым языкам не быть достойными их иметь?

Разработка программных продуктов для малоресурсных языков, в частности для языков народов России, — перспективное, интересующее сейчас многих направление компьютерной лингвистики. К таким продуктам, которые значительно облегчают изучение и исследование, а стало быть и сохранение языков, можно отнести клавиатурные раскладки, языковые корпуса, голосовые анализаторы, онлайн-переводчики и т. п.

Мы задались целью создать онлайн-переводчик для вепсского языка (переводящий с русского на вепсский и с вепсского на русский).

Обычно для самообучения переводящей программы используется большой объём текстов (в особенности параллельных, то есть в нашем случае текстов на вепсском и русском языках с одним и тем же содержанием), но в случае с вепсским языком большого объёма текстов не было. Кроме того, проблемой с точки зрения обучения переводчика является то, что доступные тексты на вепсском языке представлены не только в нормативном виде, но и в диалектном (у вепсского языка три отличающихся друг от друга диалекта: северный, средний и южный; есть отличия и между говорами в рамках одного диалекта), да и письменная норма у вепсского языка на самом деле не одна.

В итоге в качестве формы, на которой работает переводчик, была выбрана наиболее распространённая петрозаводская вепсская письменная норма, но при переводе с вепсского на русский диалектные фразы тоже могут распознаваться правильно.

Задача по созданию онлайн-переводчика гораздо в большей степени программистская, чем лингвистическая; лингвист здесь только консультирует программиста, оценивает результаты перевода. Так, в нашем случае понадобилось нарастить массив текстов, на которых переводчик в автоматическом режиме должен был обучить сам себя, и для этого было принято хитроумное решение взять массив финских и эстонских текстов и придать им «псевдовепсский» вид. Так, для преобразования финского в «псевдовепсский» мы прописали автоматическую переделку дифтонга yö в гласный ö (так из фин. työ можно получить вепс. tö ‘работа’), дифтонга ie — в гласный e, i на конце слова — в знак смягчения ’ (фин. kieli → вепс. kel’ ‘язык’), ista и istä на конце слова – в išpäi (фин. kodista – kodišpäi ‘из дома’) и т. д., автоматически переделали и некоторые служебные слова, а именно формы отрицательного глагола и союзы (фин. emme → вепс. em ‘мы не’, фин. jos → вепс. ku ‘если’, фин. että и jotta → вепс. miše ‘что’). Естественно, не во всех случаях такие преобразования дают верные результаты (например, -ista правильно переделается в -išpäi, только если речь идёт о формах элатива, а -lla верно изменится в -lda только в инфинитивах по типу tulla → tulda ‘прийти’, но не в формах адессива), да и вообще в финском и эстонском языках есть довольно много лексических основ, которых в вепсском нет, но, тем не менее, полученные таким образом «синтетические» корпуса — один «полуфинский-полувепсский», другой «полуэстонский-полувепсский» — сыграли свою техническую роль, благодаря чему переводчик стал хорошо справляться с настоящим вепсским языком.

Так как самообучение переводящей программы происходило постепенно и требовало новых данных, на финальном этапе в базу переводчика был добавлен фрагмент сделанного мной вручную перевода на вепсский язык сказки Г. Х. Андерсена «Снежная королева» (при этом значительную часть текстов, взятых в базу данных переводчика на начальном этапе, также составили мои переводы — переводы рецептов для книги «Мамина кухня», переводы историй для книги «Сказания вепсского подворья», перевод книги «Маленький принц» и др.). В итоге мы получили онлайн-переводчик, который работает в обе стороны на довольно высоком уровне, может повысить престиж вепсского языка в обществе и интерес к нему. Он призван явиться эффективным вспомогательным инструментом при изучении вепсского языка и работе с ним для школьников, студентов, журналистов, преподавателей и всех интересующихся вепсским словом.

Алексей Куташов,
ML-энтузиаст

К этому проекту я решил подключиться потому, что языки с ограниченными ресурсами – это всегда вызов. На форумах на запросы о включении редких языков в программные продукты почти всегда отвечают отказом, потому что требуется много работы в ходе обучения, но пользователей у конечного продукта будет совсем мало. При этом в случае с малыми языками нет гарантированного результата, а конечный продукт во многих случаях вряд ли будет использоваться на практике.

Для меня это было как возможностью приобрести ценный опыт работы над интересным проектом, так и реализацией стремления помочь людям лучше понимать друг друга. Меня очень удивила активность, тот объём труда, который сообщество энтузиастов вепсского языка готово вложить, чтобы сохранить и развить то, что они так ценят.

В ходе проекта мы успешно обучили переводческие модели между вепсским и русским языками. При этом особенностью было то, что исходно у нас было крайне мало параллельных текстов (всего около 3 тысяч предложений), да и просто предложений на вепсском (без перевода) у нас тоже было мало — всего около 125 тысяч. Поэтому для расширения обучающей базы переводчика мы использовали пару интересных, нетривиальных приемов.

Основные технические особенности нашей работы состояли в следующем:

1. Особенности архитектуры. Мы использовали глубокую модель трансформеров (по 20 слоев на кодер и декодер) с широким слоем FF=9216 (эффективный размер), функцию активации GeGLU (gated gelu) и относительное позиционное кодирование. Словарь использовался общий, SPM с активной опцией byte_fallback. Итого — 457M параметров.

2. Особенности датасета. Всего у нас было 3 тысячи параллельных предложений и 125 тысяч предложений, переведенных с вепсского языка на русский при помощи API Gemini Pro 1.5 (как выяснилось, качество такого перевода, особенно при учёте контекста, оказалось вполне приемлемым; видимо, сказывается близость вепсского языка к финскому и эстонскому).

3. Модель мультиязычная и обучалась с нуля. При этом мы использовали методы Transfer Learning и Back-Translation следующим образом:

- так как вепсский язык относится к группе финно-угорских языков и похож на финский и эстонский, мы добавили тексты на этих языках в датасет переводчика;

- также мы разработали алгоритм, который должен был максимально сблизить облик финских и эстонских текстов с текстами на вепсском, — таким образом мы создали дополнительные датасеты с новыми «синтетическими языками»;

- мы обучили модели VEP_RU (переводящую с вепсского на русский) и RU_VEP (переводящую с русского на вепсский) сначала на «технических» данных (на текстах из финского, эстонского датасетов и их синтетических аналогов, в каждом из них было примерно по 10 млн пар предложений), а уже потом совершили тонкую настройку, используя массив данных непосредственно на вепсском языке;

- используя модель RU_VEP, мы создали дополнительный датасет из предложений, переведённых с исходного русского на вепсский (ок. 2 млн пар) и дополнительно дообучили на этих данных модель VEP_RU, получив вследствие этого уже довольно хороший результат;

- используя эту новую модель VEP_RU, мы снова перевели на русский язык имеющиеся у нас 125 тысяч вепсских предложений в дополнение к уже имевшимся переводам, сделанным ранее при помощью API Gemini Pro 1.5, и дообучили финальную модель RU_VEP на полученных таким способом данных.

Стоит добавить, что создание онлайн-переводчика теоретически открыло возможность использовать его базу данных в целях дальнейшей цифровизации вепсского языка. Теперь любая крупная компания, используя эту базу, может создать как свои продукты на этой основе (свои переводчики, чат-боты, модели определения языка, модели эмбеддингов (для категоризации, суммаризации текста и т.п.)), так и переводчики между вепсским и всеми остальными основными языками мира.

Сами модели и датасет можно найти на странице переводчика.