О публикации результатов полевых лингвистических исследований

Константин Поливанов Олег Сериков
{k.polivanov, serikov}@iling-ran.ru

В последние годы цифровые ресурсы, создаваемые полевыми лингвистами, получают всё больше внимания. Сами исследователи прибегают к цифровым ресурсам для получения оперативного доступа к собранным и обработанным ранее данным. Вместе с тем, другие исследовательские коллективы получают доступ к материалам коллег, что позволяет избегать дублирования гипотез и извлечь максимальную пользу из наработок всего сообщества. Носители языков нередко прибегают [Bjerva J. et al., 2020; Gorman K. et al., 2020] к использованию ресурсов, например, в целях консультации или поддержки преподавания. В конце концов, исследователи в области искусственного интеллекта вовлекают [Conneau A. et al. 2020; Feng F. et al. 2020] данные всё большего количества языков к анализу, построению и оценке языковых систем ИИ.

Ранее исследователи часто прибегали к построению корпусных ресурсов с чистого листа, а в последние годы выделились несколько систем, в рамках которых существенно проще выполнить аналогичную задачу. В результате, «молодые» ресурсы оказались существенно лучше унифицированы, сейчас с ними проще работать. Более ранние ресурсы, наоборот, хоть и накопили более существенный опыт взаимодействия с пользователями, но оказываются многократно более сложны в размещении и поддержке. Для реализации централизованной поддержки обоих «поколений» корпусных платформ, была разработана гибкая система. позволяющая разместить в едином портале различные корпуса.

В рамках базовой технологии для публикации новых данных используется платформа Tsakorpus [Arkhangelskiy T., 2019]. Tsakorpus активно поддерживается разработчиками, и предлагает связки с популярными у полевых исследователей инструментами (ELAN, FieldWorks), что упрощает перенос данных от исследователей в платформу. Вместе с тем, существуют пока нерешенные технические задачи, такие, как, например, осуществление вышеупомянутого переноса без привлечения технических специалистов для запуска скриптов командной строки. Подключение сторонних или реализованных на других платформах ресурсов реализовано при помощи механизма Iframe. В итоге получается многогранная техническая конструкция, представляемая пользователю в едином интерфейсе, в рамках сайта corpora.iling-ran.ru .

Совместно с коллегами ведется активная работа по подготовке и загрузке новых данных. Для поддержания актуальности опубликованных данных была разработана методология пополнения выложенных корпусов.

В докладе мы опишем опыт Лаборатории по подготовке, публикации и сопутствующей поддержке онлайн-версий корпусов, разрабатываемых нашими коллегами из России (Ительменский, Кетский, Горномарийский, Эвенкийский, Куллуи языки) и за рубежом (Цахурский, Татский, Северный Талышский, Рутульский, Мегрельский, Лезгинский, Крызский, Каратинский, Даргинский, Восточно-Армянский). Мы представим сложившуюся методологию по ведению публикационной работы, и опишем сложившийся стек технологий, с описанием технических и практических достоинств и недостатков сложившейся картины.

Литература

Bjerva J. et al., 2020 — Sigtyp 2020 Shared task: Prediction of typological features //arXiv preprint arXiv:2010.08246. – 2020.

Gorman K. et al., 2020 — The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion //Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. – 2020. – С. 40-50.

Conneau A. et al., 2019 — Unsupervised cross-lingual representation learning at scale //arXiv preprint arXiv:1911.02116. – 2019.

Feng F. et al., 2020 — Language-agnostic bert sentence embedding //arXiv preprint arXiv:2007.01852. – 2020.

Arkhangelskiy T., 2019 — Corpora of social media in minority Uralic languages //Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. – 2019. – С. 125-140.