Рецензия на проект Дарьи Савиной «Автоматическое глоссирование малоресурсного языка» (на нивхском материале)

Леонид Зайцев

Данный проект мне кажется полезным и востребованным как для лингвистики, так и для языкового сообщества, поскольку ставит задачу автоматизации важной части работы по документации и, при оптимистичном взгляде на вещи, ревитализации языка. Особенно интересно решение этой задачи при ограничениях, которые задаёт нивхский материал, не имеющий ни достаточно большого корпуса для обучения собственной большой языковой модели, ни близких родственников, на которые в этом отношении можно было бы надеяться. Автор проекта создаёт решение, отвечающее и современному развитию компьютерной лингвистики, и специфике языкового материала, сочетая запросы к большой языковой модели с алгоритмами поиска выражений, похожих на целевое, что особенно полезно ввиду сложной системы сегментных чередований, действие которых даёт по несколько обличий для одной и той же морфемы.

В числе сильных сторон работы можно назвать внимательное отношение к языковой структуре, например, в том, что касается морфосинтаксиса нивхских прилагательных, или в выделении среди имён отглагольных номинализаций. Другие важные достоинства проекта - это использование новых метрик, более подходящих для такой задачи, благодаря балансу между полнотой и точностью, специально написанный алгоритм поиска похожих слов на основе такой метрики, эффективное использование ограниченного корпуса.

На мой взгляд, в исследовании не хватает более детализированного решения проблемы алломорфии в связи с чередованиями, которое, возможно, привело бы к ещё большему сокращению числа ошибок модели. Также на результаты могли влиять не в лучшую сторону проблемы уже существующей корпусной разметки, как формальные, так и содержательные (вплоть до глубоко теоретических), которые, конечно, не имеют отношения к работе автора проекта, напротив, вклад таких проблем автором учитывается при анализе ошибок модели.