ОБРАТНЫЙ ПЕРЕВОД КАК СПОСОБ РАСШИРЕНИЯ РУССКО-УЗБЕКСКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА

Анварова Мафтунахон Дилмурод Кизи

Аннотация. *Рассматривается использование обратного перевода для получения синтетических параллельных данных. Определены условия отбора узбекских текстов, контроля смысловой эквивалентности и включения автоматически созданных пар в обучение русско-узбекской системы. Обоснована необходимость раздельного учёта происхождения данных и независимой проверки их эффективности.*

Ключевые слова: *обратный перевод, синтетические данные, параллельный корпус, русский язык, узбекский язык, нейронный машинный перевод.*

Недостаток качественных параллельных текстов ограничивает возможности обучения русско-узбекских систем машинного перевода. Исследование машинного перевода тюркских языков, проведённое Дж. Мирзахаловым и соавторами, показывает значимость ресурсной обеспеченности и специально подготовленных тестовых материалов.[38] Одним из способов расширения обучающих данных является обратный перевод. Цель настоящего обзора — определить условия его применения к русско-узбекскому направлению с учётом лингвистического качества создаваемых пар. Используется сравнительное обобщение научных публикаций.

В нейронном машинном переводе обратный перевод предполагает автоматическое преобразование монолингвальных текстов целевого языка в исходный язык. Полученные пары дополняют корпус для обучения основного направления. Такой способ использования монолингвальных данных исследован Р. Сеннрихом, Б. Хэддоу и А. Бёрч.[39] Для системы, переводящей с русского на узбекский, исходным материалом расширения служат узбекские тексты: вспомогательная модель переводит их на русский

[38] Mirzakhalov J. et al. A Large-Scale Study of Machine Translation in Turkic Languages. 2021. P. 5876–5890.
[39] Sennrich R., Haddow B., Birch A. Improving Neural Machine Translation Models with Monolingual Data. 2016. Разд. 3.2.

язык. При последующем обучении русский синтетический сегмент становится входом, а исходный узбекский текст — целевым образцом.

Преимущество данной процедуры заключается в сохранении естественного текста на целевой стороне обучающей пары. Узбекский материал может содержать разнообразные словоформы, устойчивые сочетания и конструкции, редко встречающиеся в имеющемся двуязычном корпусе. Вместе с тем монолингвальный текст требует предварительной проверки: происхождение из открытого источника не гарантирует нормативности и отсутствия машинной генерации. Отбор следует соотносить с задачей системы, например с переводом учебных или научных материалов, и сопровождать фиксацией источников и жанров.

Главное ограничение связано с тем, что естественность узбекского предложения не обеспечивает эквивалентности автоматически созданной русской версии. Рассмотрим иллюстративный пример: «Talaba topshiriqni hali bajarmadi» — «Студент ещё не выполнил задание». Если вспомогательная модель сформирует вариант «Студент уже выполнил задание», в обучающей паре возникнет противоречие между утвердительным и отрицательным содержанием. Этот пример сконструирован для пояснения риска, а не получен при тестировании определённой системы. Подобные ошибки особенно существенны при передаче отрицания, времени, количества и отношений между участниками ситуации.

С. Едунов и соавторы установили, что результативность обратного перевода зависит от способа генерации синтетических предложений; преимущества отдельных методов различаются в зависимости от ресурсных условий.[40] Поэтому нельзя заранее признать оптимальным один способ декодирования или единую долю синтетических пар для русско-узбекского корпуса. Вариативность формулировок может быть полезна для обучения, однако её необходимо отличать от искажения содержания. Выбор параметров должен опираться на валидационную выборку, а не на показатели итогового теста.

Для русско-узбекского материала предлагается последовательная проверка синтетических данных. Сначала исключаются пустые, дублирующиеся и ошибочно определённые по языку сегменты. Затем сопоставляются числа, имена собственные, термины и смысловые отношения. Автоматические признаки используются для выделения

[40] Edunov S., Ott M., Auli M., Grangier D. Understanding Back-Translation at Scale. 2018. P. 489–500.

сомнительных пар, после чего проводится выборочная экспертная проверка. Каждая пара должна сопровождаться указанием синтетического происхождения, версии вспомогательной модели и настроек генерации. Это позволяет выявлять источники повторяющихся ошибок и при необходимости исключать соответствующие группы данных.

Эффективность расширения корпуса целесообразно проверять сравнением обучения на исходных данных, на исходных данных с нефильтрованными синтетическими парами и с проверенными парами. Архитектура, вычислительный бюджет и тестовый набор должны быть сопоставимыми; пересечения теста с монолингвальным материалом расширения исключаются заранее. М. Пост показал зависимость BLEU от параметров обработки текста, что требует единообразного расчёта метрики. [41] Количественное сравнение следует дополнять анализом смысловых и грамматических ошибок, особенно связанных с узбекскими аффиксами и русскими предложно-падежными конструкциями.

Таким образом, обратный перевод представляет способ контролируемого расширения обучающего корпуса, при котором естественные узбекские тексты соединяются с автоматически полученными русскими соответствиями. Его практическая ценность определяется сохранением смысла, соответствием предметной области и прозрачностью происхождения данных. Для русско-узбекской системы обоснованным направлением дальнейшего исследования является экспериментальное сравнение способов генерации и фильтрации синтетических пар на независимом материале.

Список литературы

1. Mirzakhalov J., Babu A., Ataman D. et al. A Large-Scale Study of Machine Translation in Turkic Languages // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. P. 5876–5890. DOI: 10.18653/v1/2021.emnlp-main.475.

2. Sennrich R., Haddow B., Birch A. Improving Neural Machine Translation Models with Monolingual Data // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Vol. 1: Long Papers. Berlin: Association for Computational Linguistics, 2016. P. 86–96. DOI: 10.18653/v1/P16-1009.

[41] Post M. A Call for Clarity in Reporting BLEU Scores. 2018. P. 186–191.

3. Edunov S., Ott M., Auli M., Grangier D. Understanding Back-Translation at Scale // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. P. 489–500. DOI: 10.18653/v1/D18-1045.

4. Post M. A Call for Clarity in Reporting BLEU Scores // Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels: Association for Computational Linguistics, 2018. P. 186–191. DOI: 10.18653/v1/W18-6319.