**Проблемы машинного перевода малоресурсных языков на примере тайского языка**

Парфенова Светлана Сергеевна

Студентка Южно-Уральского государственного университета, Челябинск, Россия

В эпоху глобализации все более актуальным становится сохранение языкового разнообразия. Современные машинные переводчики работают с сотнями языков. Однако большинству существующих языков не хватает цифровых ресурсов для эффективной обработки и анализа. Носители малоресурсных языков (далее – МЯ) могут сталкиваться с такими сложностями, как отсутствие систем машинного перевода и распознавания текста или речи. Здесь будут рассмотрены проблемы машинного перевода МЯ на примере тайского языка.

Малоресурсными можно назвать те языки, у которых недостаточно данных для обучения (корпусов, словарей и грамматик), а также инструментов для обработки языка. Среди методов, используемых для работы с МЯ, можно отметить трансферное обучение, создание синтетических корпусов при помощи обратного перевода и обучение без учителя [Zhong et al. 2024].

Тайский можно отнести к МЯ по сравнению с более распространёнными, английским, китайским или испанским. Так, одни из самых крупных корпусов тайского языка –  Thai web corpus (695 млн слов), Thai National Corpus (33 млн слов), HSE Thai Corpus (50 млн слов). В то же время Оксфордский корпус английского языка содержит около 2,1 млрд слов. Некоторые из существующих корпусов аннотированы. Например, Thai National Corpus, Thai Nested Named Entity Recognition Corpus и др. Библиотеки, такие как PyThaiNLP, DeepCut и WordCut, предлагают инструменты для токенизации, анализа частей речи и поиска именованных сущностей в тайском языке.

Помимо дефицита цифровых ресурсов, задачи, связанные с обработкой, усложняются из-за особенностей структуры тайского. Будучи изолирующим языком, он требует высокого уровня анализа контекста при обработке сразу по нескольким причинам:

* Отсутствие словоизменительной аффиксации. Передача грамматических значений (число, род, вид и пр.) в основном ложится на служебные слова.
* Наличие глагольной сериализации. В сериальной конструкции последовательность из нескольких глаголов функционирует как единый предикат. При этом в разных языках последовательность глаголов или значение подобных конструкций может отличаться, что затрудняет трансферное обучение [Морев 1991: 116].

В тайском языке выделяют 3 регистровых и 2 контурных тона. Это представляет определенную сложность для систем синтеза и распознавания речи, так как требует точного анализа высоты звука для смыслоразличения в рамках слов [Морев 1991: 41].

Сегментация предложений -- распространенная проблема при обработке малоресурсных языков. Тайский, как язык scriptio continua, использует пробелы не для разграничения слов, а для обозначения окончания предложения или клаузы. Задача сегментации облегчается тем, что большинство тайских слов состоит из 1-2 слогов, однако заимствованные слова часто бывают многосложными. Для решения этой проблемы используются различные подходы: предварительное обучение на размеченных данных, расчет частотности последовательностей букв, поиск наиболее длинного совпадения в словаре и метод условного случайного поля, чтобы определить является ли символ начальным [Poncelas et al. 2020].

Для тайского, как и для многих МЯ характерна сложная система средств выражения вежливости. Помимо частиц и местоимений, выражающих отношение к собеседнику, в тайском языке есть отдельные лексические системы, относящиеся к религии и королевской семье [Морев 1991: 66].

Так как нейронные системы требуют большого количества данных, нехватка цифровых ресурсов заметно сказывается на качестве машинного перевода. На данный момент тайский язык поддерживают такие системы, как Google Переводчик, Яндекс Переводчик, Microsoft Translator, DeepL, iTranslate и Papago. Для оценки качества перевода был выбран Google Переводчик, как самый популярная и одна из самых развитых систем машинного перевода.

Исследование проводилось на базе параллельного корпуса TedTalks на русском и тайском языках из проекта OPUS. Этот корпус был выбран, поскольку он состоит из текстов в публицистическом стиле и, следовательно, отличается разнообразием языковых средств. Корпус на тайском языке был переведен на русский при помощи Google Переводчика и сопоставлен с русскоязычным корпусом с помощью алгоритма BLEU. Кроме того, была проведена автоматическая оценка лингвистических характеристик машинного и ручного перевода.

|  | Ручной перевод | Машинный перевод |
| --- | --- | --- |
| Количество предложений | 453 | 416 |
| Количество слов | 6016 | 6480 |
| Части речи | Прилагательные: 462;  Имена собственные: 4854;  Глаголы: 369;  Существительные: 219;  Числительные: 84;  Союзы: 28 | Прилагательные: 430;  Имена собственные: 5437;  Глаголы: 366;  Существительные: 160;  Числительные: 69;  Союзы: 18 |
| Знаки препинания | Точка: 453;  Запятая: 619;  Двоеточие: 30 | Точка: 416;  Запятая: 725;  Двоеточие: 15 |
| Средняя длина слова | 4,87 | 4,86 |
| Средняя длина предложения | 15,81 | 18,66 |
| Количество уникальных слов | 3060 | 3143 |

Таким образом, можно сделать вывод о том, что Google Переводчик разделяет текст на меньшее количество предложений, чем это делает человек. Это может привести к дополнительным ошибкам и усложнить восприятие текста. Полученный результат BLEU составил 0,2 по шкале от 0 до 1. При этом тексты с высоким качеством перевода, как правило, набирают более 0,5.

За последнее время наблюдается стабильное развитие машинного перевода, в том числе и МЯ. Проблема обработки МЯ не имеет простых решений и однозначных ответов. Дальнейшие исследования в этой области могут поспособствовать общей эволюции машинного перевода.

Литература

Морев Л. Н. Сопоставительная грамматика тайских языков. М., 1991.

Poncelas A. Multiple Segmentations of Thai Sentences for Neural Machine Translation: <https://aclanthology.org/2020.sltu-1.33/>

Zhong T. Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research: <https://arxiv.org/html/2412.04497v2>