

Решение по переводу PROMT для TripAdvisor

О компании TripAdvisor

TripAdvisor® — крупнейший в мире туристический сайт, дающий возможность запланировать и совершить идеальное путешествие. TripAdvisor предоставляет проверенные рекомендации от настоящих путешественников, а также многочисленные варианты для планирования поездок с прямыми ссылками для бронирования. Сайты, работающие под брендом TripAdvisor, составляют самое многочисленное сообщество такого рода в мире: они насчитывают более 60 миллионов уникальных посетителей в месяц и более 75 миллионов отзывов и мнений.

Цель проекта

Основной контент сайта TripAdvisor составляют миллионы отзывов пользователей со всего мира. Но тем, кто собирается в путешествие, тексты должны быть доступны на родном языке. Поэтому для локализованных сайтов компания делает перевод пользовательских комментариев. Обычный ручной перевод в этом случае затруднителен из-за огромного объема контента TripAdvisor. Для решения этой проблемы TripAdvisor обращается к машинному переводу.

Российский туристический рынок с его большим потенциалом не мог пройти мимо внимания сервиса. И для создания русскоязычной версии сайта TripAdvisor решил перевести все англоязычные отзывы пользователей на русский. В качестве партнера была выбрана компания PROMT как ведущий поставщик решений по машинному переводу с русского и на русский.

Исходные требования

Решение по переводу для TripAdvisor должно удовлетворять трем основным требованиям:

1. **Перевод достаточно высокого качества, не требующий дальнейшего редактирования для чтения и понимания.**

Машинный перевод пользовательских отзывов должен быть четким и понятным. В связи с исключительным количеством пользовательских текстов редактирование машинного перевода для каждого из них невозможно. Однако при переводе содержимого TripAdvisor возникают объективные трудности:

- Контент сложен для машинного перевода, так как тексты, написанные пользователями, часто очень эмоциональны и содержат значительное количество опечаток, орфографических, стилистических и пунктуационных ошибок. Это влияет на качество автоматического перевода на решающем первом шаге исходной обработки текста.
- Другая проблема состоит в том, что в начале проекта TripAdvisor не имел достаточно больших и релевантных параллельных текстов, что затрудняло настройку решения под задачу.

2. **Автоматическая оценка качества перевода**

Оценка качества перевода — очень важный элемент любого решения, использующего машинный перевод. Так как ручная оценка всего переводимого контента TripAdvisor невозможна, решение по машинному переводу должно предоставить автоматический механизм оценки качества переведенных текстов — confidence score, чтобы дать возможность TripAdvisor публиковать переведенные отзывы только высокого качества.

3. Интеграция в производственный процесс сайта TripAdvisor

Разработчики TripAdvisor хотели бы использовать машинный перевод как веб-сервис на удаленном сервере, который получает запросы и возвращает перевод в удобном для них формате, требующем минимальной постобработки.

Развертывание решения

В мае 2011 был сделан пробный перевод партии наиболее востребованных отзывов без настроек на контент. Система даже в базовом варианте продемонстрировала достаточное качество перевода, после чего началось развертывание решения с глубокой настройкой на контент сайта TripAdvisor.

В ноябре 2011 решение PROMT было полностью интегрировано в производственный процесс TripAdvisor и с тех пор отзывы пользователей переводятся еженедельно.

Сервер перевода размещен в дата-центре PROMT и его работа полностью обслуживается специалистами PROMT.

TripAdvisor регулярно предоставляет PROMT новые лингвистические данные для дополнительной настройки, и качество перевода продолжает улучшаться.

Компоненты решения PROMT для TripAdvisor

Основные составляющие уникальной разработки PROMT для TripAdvisor:

- [PROMT Translation Server 9.5 DE](#) — надежное, производительное и масштабируемое серверное решение, позволяющее переводить большие объемы текстов.
- [PROMT DeepHybrid](#) — технология, позволяющая получить более качественный и понятный конечному читателю перевод. Для ее применения потребовались:
 - **Специальные словари.** Для данного проекта был значительно переработан специализированный словарь PROMT «Путешествия» и составлен отдельный словарь для клиента TripAdvisor. Также был разработан препроцессор ошибок в исходном тексте, т.к. в пользовательских текстах неизбежно встречаются ошибки, опечатки, сокращения и т. д., и система перевода должна их правильно распознавать, чтобы перевод передавал смысл оригинала.
 - **Языковая модель:** PROMT собрал и обработал корпус сгенерированных российскими пользователями текстов по теме «Путешествия» и использовал его для создания **языковой модели**, которая необходима для статистического постредактирования переведенных отзывов и для автоматической оценки качества перевода.
 - **Статистическое постредактирование:** Эта функция позволила PROMT регулярно улучшать первоначально настроенную систему. Отзывы, которые наиболее часто читаются российскими пользователями, проходят через человеческое постредактирование и составляют основу **таблицы статистического постредактирования**), которое делает перевод более гладким и более похожим на человеческий.
- **Автоматическая оценка качества перевода (confidence score).** Сразу после перевода отзыва система PROMT автоматически сравнивает получающийся текст с **языковой моделью**, производя измерение того, как близко перевод находится к оригинальному корпусу русских текстов.

Для определения порога confidence score, подходящего для контента TripAdvisor, PROMT сравнил автоматические оценки с человеческой оценкой переводов, выполненных несколькими лингвистами, по большой случайной выборке отзывов.

- **Специальный веб-сервис для TripAdvisor.** Решение было развернуто на выделенном сервере PROMT. На этапе интеграции был согласован формат передачи данных, наиболее удобный для разработчиков TripAdvisor, и реализована специальная функция в API.

Заключение

Решение по переводу PROMT для TripAdvisor полностью решило задачи проекта проекта, удовлетворяет всем исходным требованиям и помогает TripAdvisor переводить большой объем отзывов на русский язык.

Результаты реализации:

- Большие текстовые объемы переводятся быстро.
- Качество перевода достаточно для полного понимания.
- Затраты на перевод всего содержания значительно меньше, чем при ручном переводе даже небольшой части отзывов.
- Используется автоматическая оценка качества перевода.
- Решение интегрировано в производственный процесс TripAdvisor с минимальными затратами по разработке и поддержке со стороны TripAdvisor.

«Сайт TripAdvisor очень популярен, благодаря пользователям, которые пишут обо всем, что интересно путешественникам: отелях, ресторанах, достопримечательностях и т.д. Поэтому критически важная для бизнеса задача – сделать отзывы пользователей доступными на различных языках для путешественников во всем мире. Из-за большого количества отзывов для решения этой проблемы мы можем использовать только высококачественный машинный перевод. Мы выбрали PROMT для перевода с английского на русский язык, и он успешно работает уже в течение года. PROMT помогает нам доносить опыт англоговорящих путешественников до нашей русскоязычной аудитории».

Лорна Уилан, старший менеджер по локализации, TripAdvisor

«Такой крупный сайт, как TripAdvisor, агрегирующий создаваемый пользователями контент, – идеальный клиент для машинного перевода, и в частности для технологии DeepHybrid. Статистический подход в чистом виде работает хорошо только на больших объемах параллельных данных, в то время как PROMT DeepHybrid производит высококачественный перевод, даже когда объемы данных ограничены. Создание системы перевода для TripAdvisor было для нас сложной и интересной задачей из-за больших объемов и спонтанного характера контента. Но мы справились, и теперь десятки миллионов русскоязычных пользователей интернета могут планировать поездки, основываясь на опыте других пользователей такого популярного и качественного ресурса, как TripAdvisor».

Юлия Епифанцева, директор по развитию PROMT