PROMT КОРПОРАТИВНЫЕ РЕШЕНИЯ PROMT 22

Истина в данных
или
Секреты обучения

ЭГО ТРАНСЛЕЙТИНГ
PROMT

EGOTECH

ГК ЭГО ТРАНСЛЕЙТИНГ

EGO TRANSLATING COMPANY

Сила Слова
в цифровом мире

Письменные переводы
всех видов документации, в соответствии с действующими отраслевыми стандартами и стандартами заказчика; копирайтинг и рерайтинг

Локализация
веб-сайтов, программного обеспечения, e-commerce, мультимедиа

EGOTECH
управление терминологией, составление и редактирование глоссариев, программные решения для повышения качества и оптимизации сроков и стоимости переводов и локализации контента

Устные переводы очно и по ВКС
сопровождение переговоров, конференций, презентаций, протокольных мероприятий, шеф-монтажей и пуско-наладки оборудования на предприятиях

Верстка и графическое оформление
материалов во всех текстовых и графических редакторах, работа с чертежами любого уровня сложности

Легализация перевода,
включая удостоверение перевода печатью Компании ЭГО Транслейтинг, нотариальное удостоверение подписи переводчика, проставление штампа Апостиль и консульскую легализацию

Корпоративное обучение
иностранным языкам, тренинги по межкультурной адаптации и коммуникации

EGOTECH

Технологический суббренд Группы Компаний ЭГО Транслейтинг

30+ лет опыта, 30+ отраслей, 1Тб+ накопленных данных

Программные решения и услуги для LSP/Производства/Лингвистики

Мультифункциональная онлайн-платформа egotech.tech

Команда профессионалов

3 фактора успеха при внедрении МП

Технология

Передовые достижения современной науки позволяют получить потрясающий результат машинного перевода

Продукт

Гарантия конфиденциальности переводимых данных, гибкая настройка под переводческие бизнес-задачи

Данные

Возможность настройки для перевода терминологии и узкоспециализированных текстов

Международное сотрудничество, службы технической поддержки, локализация документации, интернет-магазинов, сайтов, user-generated content

Качество перевода

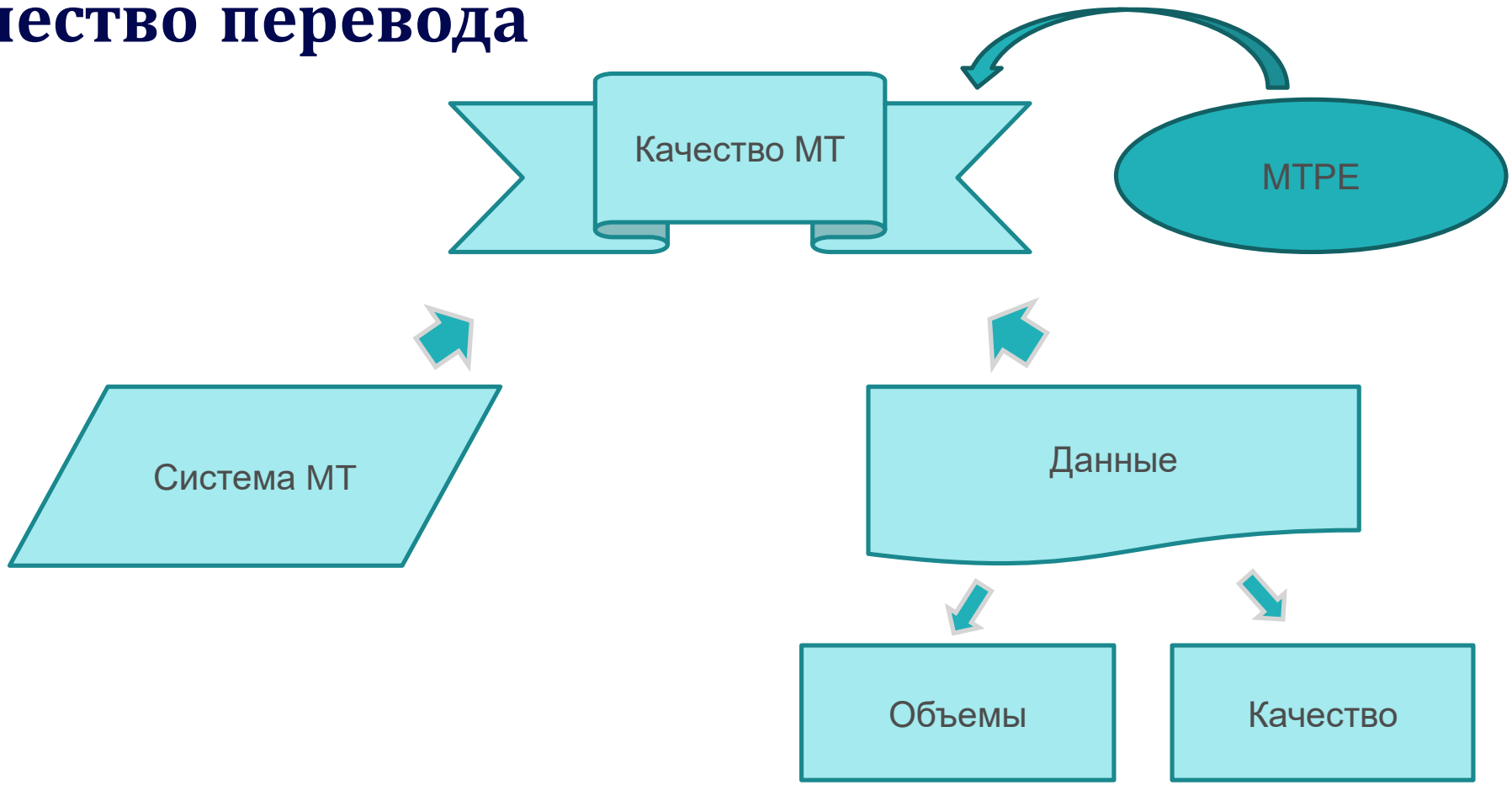
Универсальные NMT-модели
→ высокое качество
перевода общелексических текстов.

Проблема: Универсальные модели
не справляются с переводом терминологии в
узкоспециализированных текстах
и с передачей стиля.

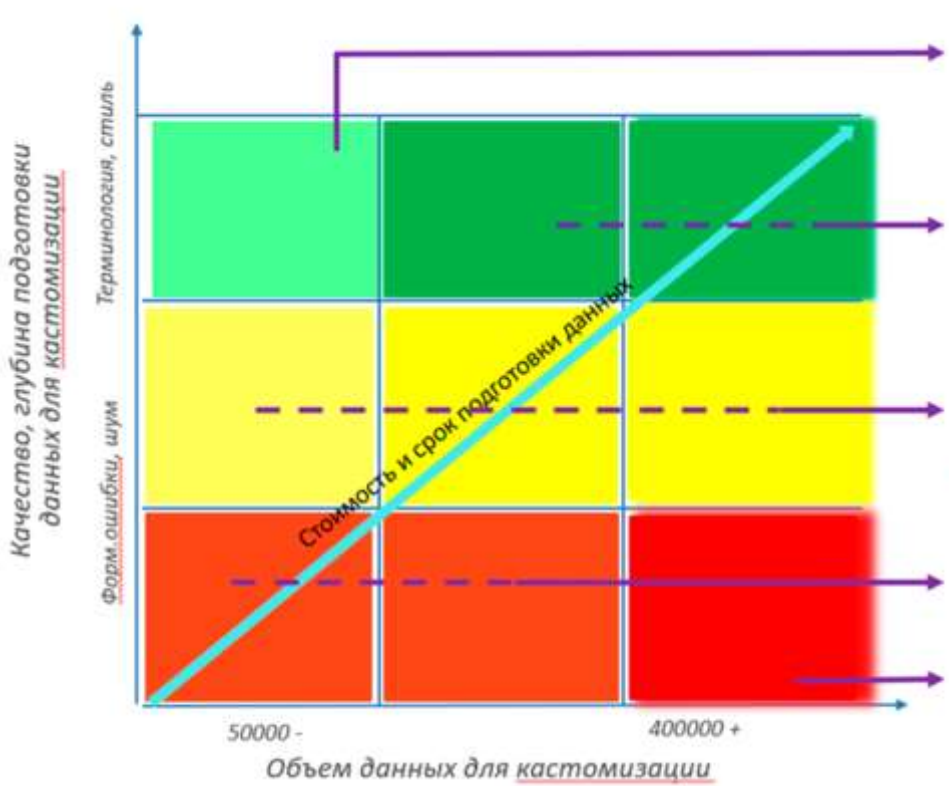
Решение : настройка NMT на параллельных
данных

Экспертиза и опыт EGOTECH по работе с данными

Качество перевода



Влияние объемов на качество



Маленькая база-мечта. Тюнинг.
Не очень долго, не очень дорого, требовательно к ресурсам.
Может сработать для решения определенных задач

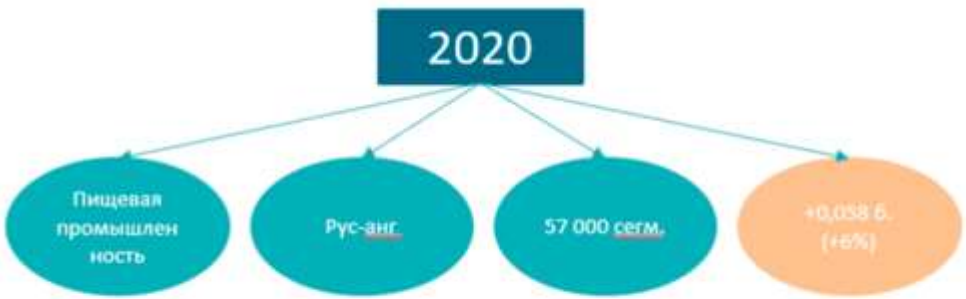
База-мечта.
Долго, дорого, требовательно к ресурсам.
Обнадеживающе по качеству.
Сомнительно по эффективности вложений.

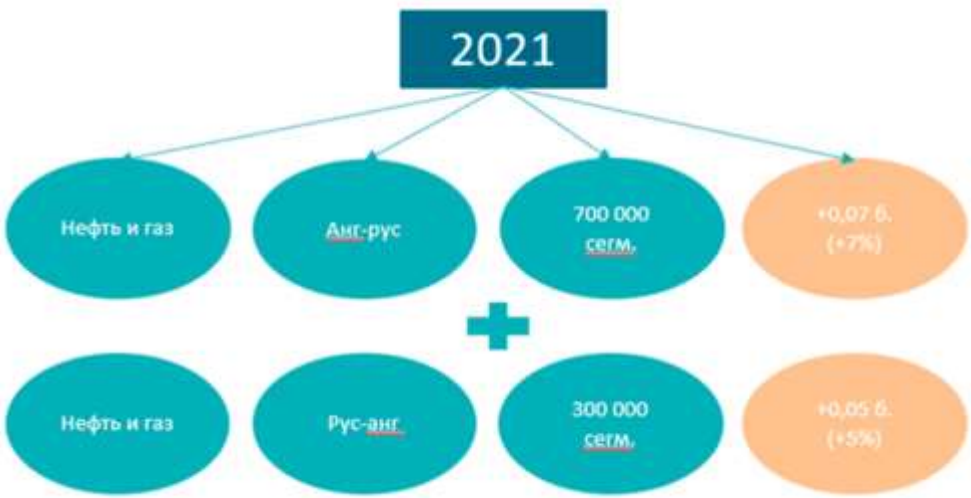
Зона вероятности.
Не очень долго, не очень дорого, не очень требовательно к ресурсам.
Труднопрогнозируемо по результату (зависит от факторов)

Быстро, дешево. Скорее всего, бессмысленно (зависит от факторов)

Быстро, дешево. Скорее всего, вредно (зависит от факторов)

Эксперименты EGOTECH по подготовке данных

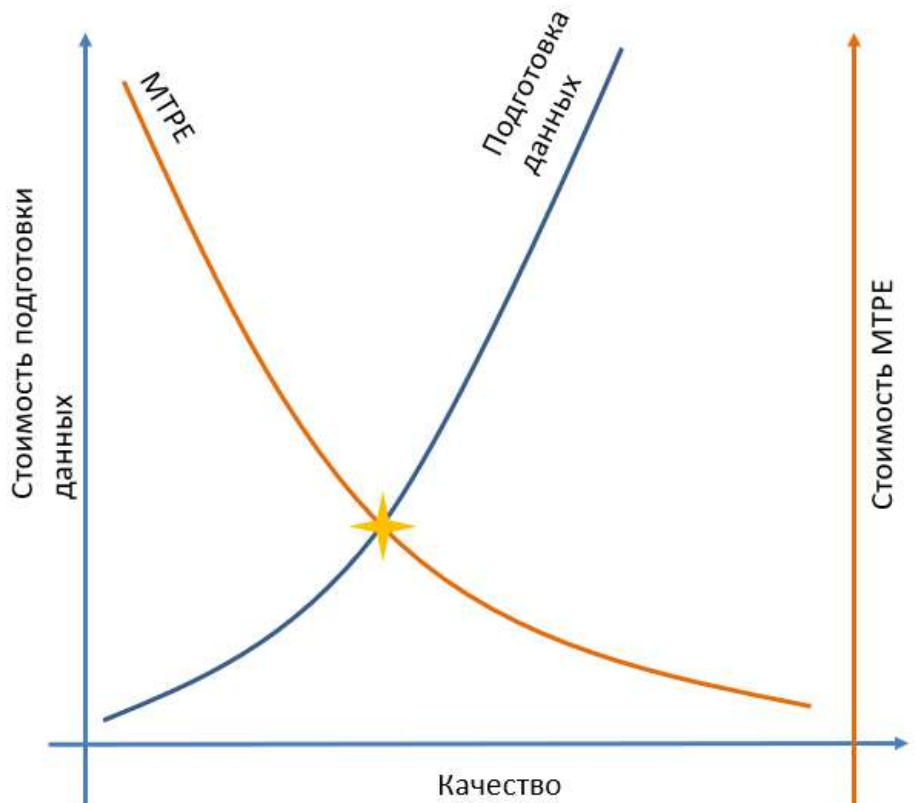




Все очень сильно зависит от:

- Языковой пары
- Тематики
- Параметров ваших метрик
- Подготовительных работ с данными

Влияние MTPE на качество



Возможности платформы EGOTECH

Возможности EGOTECH

Семейство решений EGOTECH

ET QA Check	для корректорской правки перевода
ET QA Term	для проверки терминологии в переводах и ТМ
ET TermExtract	для выборки терминов и создания одноязычного словаря
ET TMX Processor	для сокращения объема файла и «резки» и «объединения» tmx
ET TMX Cleaner	для быстрой автоматической очистки ТМ
ET TMX Clust	для создания ТМ по узким тематикам
ET TMX Editor	для редактирования tmx файлов в режиме онлайн
ET QA Metrics	для количественной оценки качества перевода

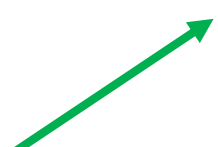
Подготовка данных для обучения систем MT

В результате обработки клиент получает нормализованный корпус, то есть очищенную БД, с помощью которой СМП могут повышать свою результативность и производительность, путем повторения итераций машинной обработки (распознавания) корпуса текста.

Обработка датасета может проводиться в несколько этапов:

- Техническая
- Корректорская
- Терминологическая

Выбор видов очистки зависит от исходного качества датасета!



!Как правило, техническая очистка является достаточной при сопоставлении параметров качества, времени на обработку и стоимости.

Инструменты PROMT для настройки MT

PROMT Training Add-on

Возможности

- Работает вместе с PROMT Neural Translation Server
- Тренировка на стороне клиента
- Данные для тренировки – файлы в формате tmx, UTF-8
- Рекомендуемый объем- от 50 000 сегментов
  (все требования на сайте)

Безопасность&конфиденциальность

- Работает офлайн
- Тренировочные данные не доступны третьим лицам

Результат

- Специализированная модель
- Модель подключается в PROMT Neural Translation Server через профиль перевода
- Профиль перевода доступен для перевода текста, документа, сайта, в CAT-инструменте (Memsource, SDL Trados)

Влияние объема данных на рост качества

Количество обучающих данных	Средний прирост BLEU score
10 000 пар предложений	+0,3
50 000 пар предложений	+3
100 000 пар предложений	+6
250 000 пар предложений	+10

PROMT Neural Dictionary

Возможности

- UI в серверных и десктопных продуктах PROMT (*Мой словарь*)
- Поддержка только существительных
- Поддержка слов и словосочетаний
- 1 ключ - 1 перевод
- Автоматический и ручной ввод

Особенности

- Метрика BLEU покажет ухудшения в качестве перевода
- Скорость перевода со словарем может незначительно снизится

Не рекомендуется для

- Частотных слов с общелексическими переводами
  accident – авария
- Частотных слов с узкоспециальным переводом
  vessel - сосуд

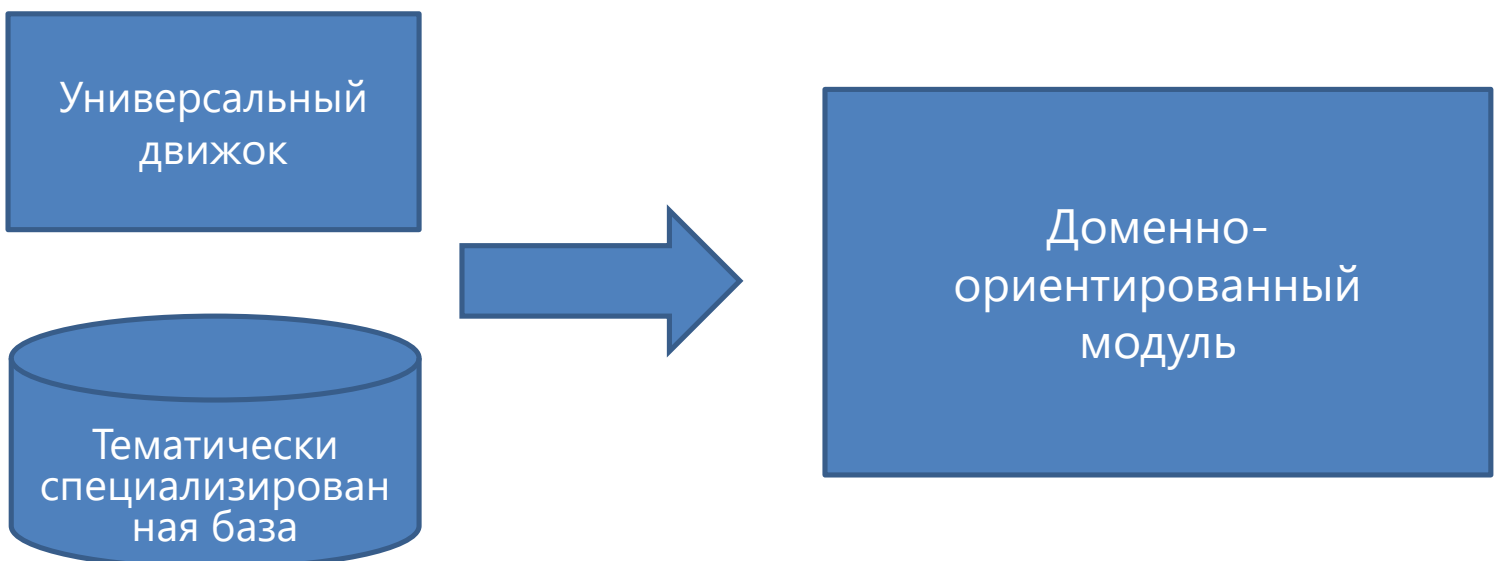
Рекомендуется для

- Имен собственных
- Аббревиатур
- Словосочетаний

Предложение PROMT & ЭГО ТРАНСЛЕЙТИНГ

Почему PROMT и Ego Translating?

Качественная технология машинного перевода PROMT + проработанная технология работы с данными EGO Translating



EGOPROMT – это доменно-адаптированные модели для решения ваших задач!

Специализированная модель «Нефть и газ»

- Показатели по метрикам:

Метрика	Показатели по стоковой модели	Показатели кастомизированной модели
BLEU	0,39	0,46
ChrF	0,37	0,47

Общая тенденция прироста высокая ≈ 10%!

- Результаты экспертного тестирования:
  - С точки зрения комплексной оценки качества перевода, модель «Нефть и газ» показала большую результативность и эффективность при переводе отраслевой терминологии, в предметной области.
  - Уменьшилось количество ошибок словоупотребления.
  - Уменьшилось число встречаемых ошибок грамматического несоответствия.
  - Общее количество ошибок, стандартных для машинного перевода, снизилось ≈ на 20%

Предложение PROMT и Ego Translating

Возможность получить специализированную модель перевода, подготовленную и натренированную на Ваших данных!

Мы предоставляем:

- Уникальную технологию машинного перевода
- Помощь с подготовкой данных

Что требуется от вас:

- ✓ Желание получить решение, способное справляться с целевой задачей оптимальным путем
- ✓ Данные для тренировки

EGOTECH

Спасибо за внимание!

Маргарита Меняйлова
+7 928 361 46 57
margarita.sehposyan@egotranslating.ru

Юлия Епифанцева
+ 7 911 2103861
julia.epiphantseva@promt.ru