Нейронный машинный перевод: на службе у профессионального переводчика

PROMT

5 мифов о МП

- МП – это онлайн-сервисы и приложения
- МП пользуются те, кто не знает иностранных языков
- Обучение системы не нужно
- МП заменит профессиональных переводчиков
- Из-за МП никто не будет изучать иностранные языки

PROMT

NEURAL MT

ХАЙП ИЛИ ЧУДО

Краткая история машинного перевода

Rule-Based MT

Машинный акцент в переводе, предсказуемый результат

ОБУЧЕНИЕЧЕРЕЗ СЛОВАРИ

С 1950 гг.

Statistical MT

Более гладкий перевод, чем в RBMT, но непредсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ

С 2000 гг.

Neural MT

Перевод, без машинного акцента, иногда непредсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ

С 2015 гг.

PROMT

Предпосылки для Neural MT

- Рост вычислительных мощностей
- Рост текстовых данных на разных языках в digital формате
- Успехи в области нейросетевых технологий в разных отраслях

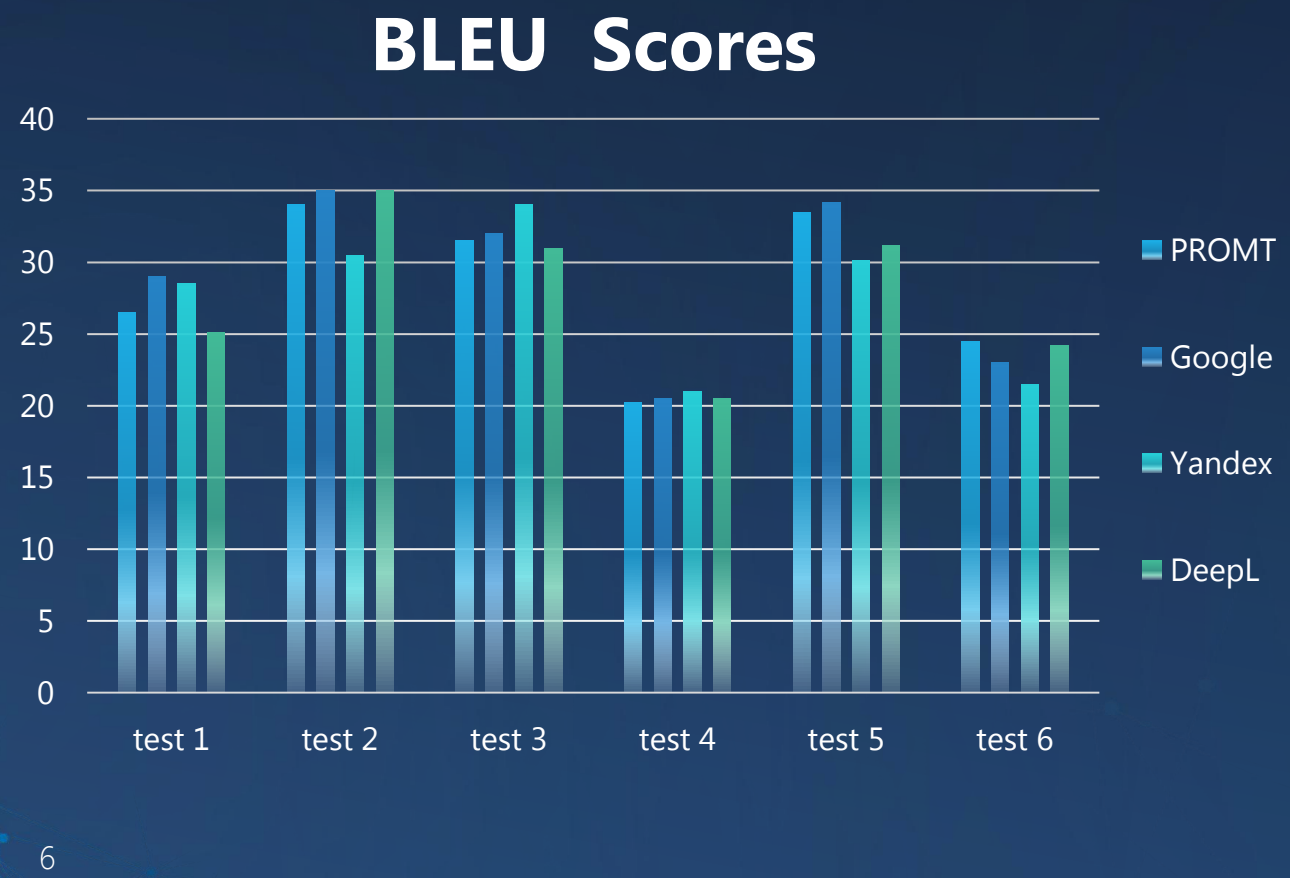
Некоторые факты о Neural MT

- Более 100 разработчиков технологии на основе open-source и собственных разработок
- Более 500 научных публикаций
- Десятки публикаций в ведущих изданиях по всему миру

PROMT

PROMT Neural, Google, Yandex, DeepL

Тестирование на текстах общей тематики, ER



6

Экспертная оценка

- перевод легко читается
- корректная структура предложения
- нет ошибок согласования
- перевод требует незначительного или вообще не требует редактирования

PROMT

Задача перевода решена?

- **Не все тексты** переводятся хорошо
- **Обучение**
  На чем тренировать NMT?
  Сколько нужно данных и т.д.?
- **Перевод документов**
  остается сложной задачей

Типичные ошибки NMT

- Перевод имен собственных
- Нарушение единства терминологии
- Нейрологизмы

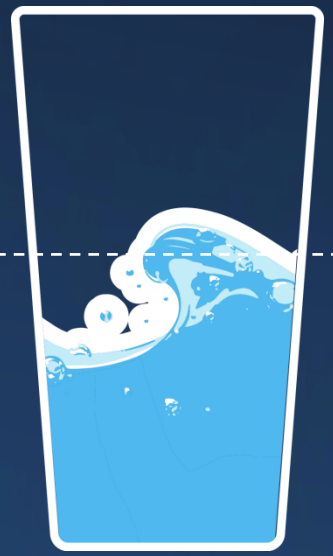
PROMT

МП & ПЕРЕВОДЧИКИ

ОПЫТ И РЕКОМЕДАЦИИ

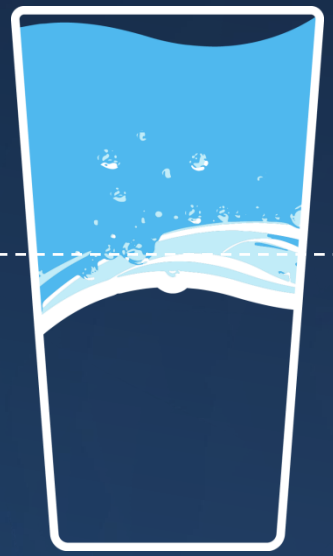
Отношение к МП

Не переводчики



Стакан наполовину полон

Переводчики



Стакан наполовину пуст

Положительный опыт с МП

- В проектах, где достаточно «базового качества»
- В проектах, где достаточно «базового качества» + light post-editing
- В проектах по локализации документации к ПО (сокращение в расходах на 15-35%)

PROMT

2 сценария использования МП

Как основа для постредактирования

Как подстрочник для справки

PROMT

Подходит ли MT для Вашего материала?

- Стартовое качество перевода
- Возможность настройки
- Поддержка исходного формата
- Возможность редактирования результата

Шаг1. Проверка стартового качества перевода

Выбрать несколько типичных фрагментов

- Короткие абзацы, а не отдельные предложения
- Избегать заголовков, названий таблиц, картинок и т.д.
- Без ярких языковых особенностей (например, узкоспециализированные аббревиатуры, формулы, имена собственные и т.д.)
- Без технических проблем (например, отсутствие разрывов строк или некачественное распознавание текста)

Перевести фрагменты с помощью МП

Шаг3. Проверка стартового качества перевода

Оценить качества МП

МЕТОД

- Если есть профессиональный перевод выбранных фрагментов, то сделать автоматическое сравнение двух переводов (редакционное расстояние)
- Если профессионального перевода нет, то выделить ошибки по типам (лексика, грамматика, орфография, форматные проблемы), определить критичность ошибок.

ВЫВОДЫ

- Редакционное расстояние меньше 30-40% - стоит использовать МП для постредактирования
- Редакционное расстояние больше 30-40% - оценить целесообразность использования МП в качестве подстрочника

PROMT

PROMT

ПРОГРАММЫ ДЛЯ МП

Способы работы

- PROMT **Neural Translation Server –** веб-интефейс
- CAT-система + плагин к PROMT **Neural Translation Server**

Перевод
в PROMT Neural Translation Server

- Перевод документов целиком с последующим редактированием
- Перевод фрагментов текстов в веб-интерфейсе
- Перевод через приложение **PROMT Агент**

Настройка PROMT NMT

Настройка на стороне PROMT

Материалы для настройки

- Релевантные параллельные данные (раннее сделанные переводы в формате tmx)
- Глоссарии с терминологией

Гибридная технология PROMT Neural

Трансформер,
Marian NMT

Алгоритмы глубокого семантико-синтаксического анализа

PROMT

Объемы данных PROMT Neural Универсальная модель, ER

Корпуса NMT

«ЧЕЛОВЕЧЕСКИЕ»

WMT, OPUS, PROMT корпуса /46,6 млн сегментов

«СИНТЕТИЧЕСКИЕ»

Wikipedia / 22 млн сегментов
Новости / 22 млн сегментов

«РАЗМЕЧЕННЫЕ»

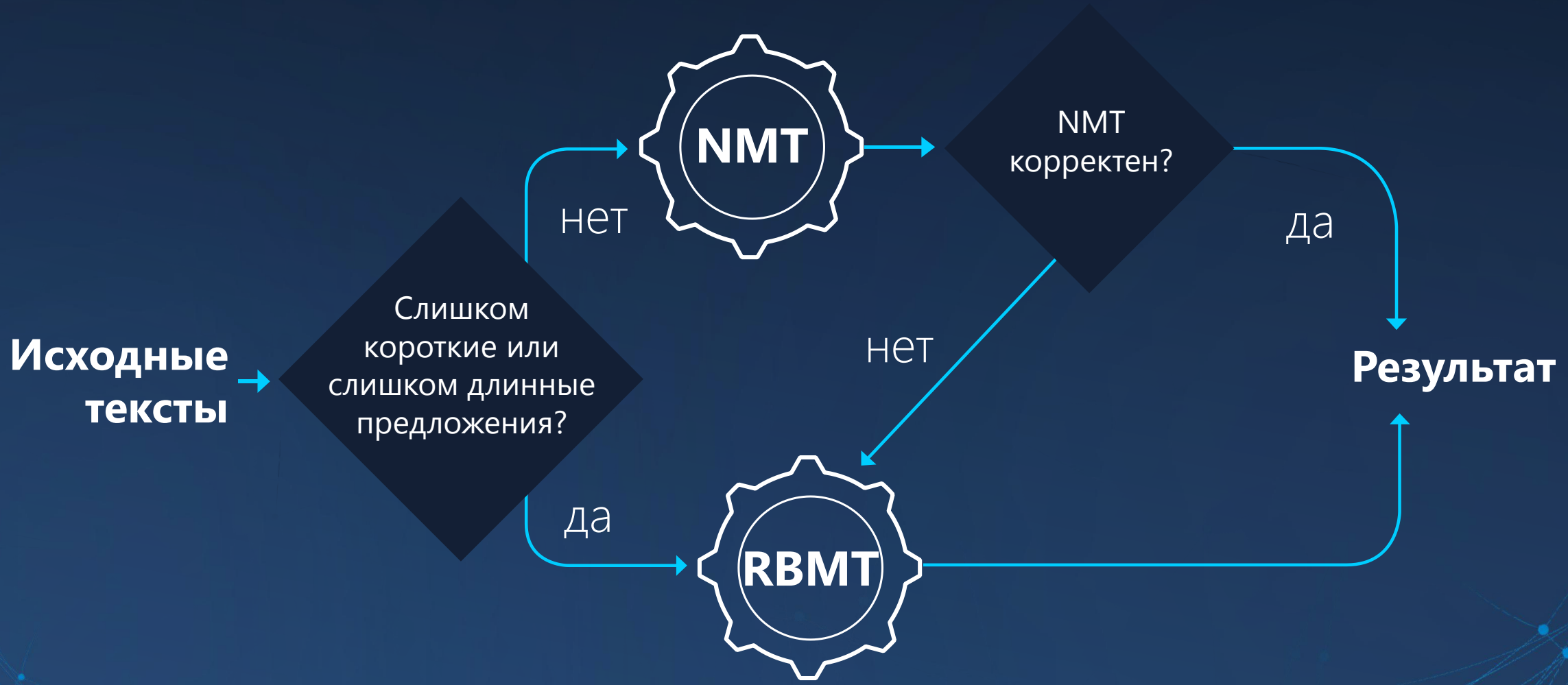
«Человеческие» и «синтетические» корпуса с разметкой для управления терминологией

Словари и алгоритмы RBMT

Двуязычные словари общей лексики для распознавания более 2 млн английских словоформ и синтеза русских словоформ + алгоритмы распознавания имен собственных и неологизмов

PROMT

Как работает PROMT Neural?



PROMT

ДАННЫЕ

ДВУЯЗЫЧНЫЕ И МОНОЯЗЫЧНЫЕ КОРПУСА И ГЛОССАРИИ ЗАКАЗЧИКА ДЛЯ ТРЕНИРОВКИ

Что такое тематически однородные данные?

Тематика: нефть/газ, энергетика ....

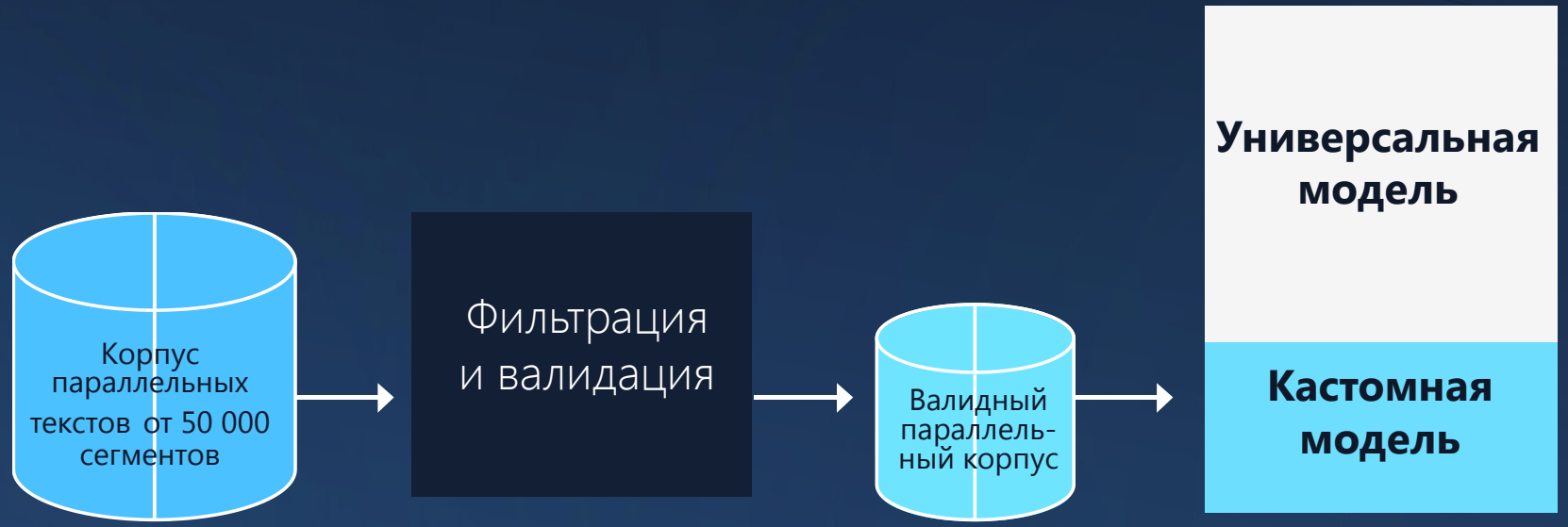
Тип текста: договоры, инструкции, User-generated Content...

Заказчик особенности текстов разных компаний

Переводчик: особенности индивидуального стиля

Тренировка PROMT Neural

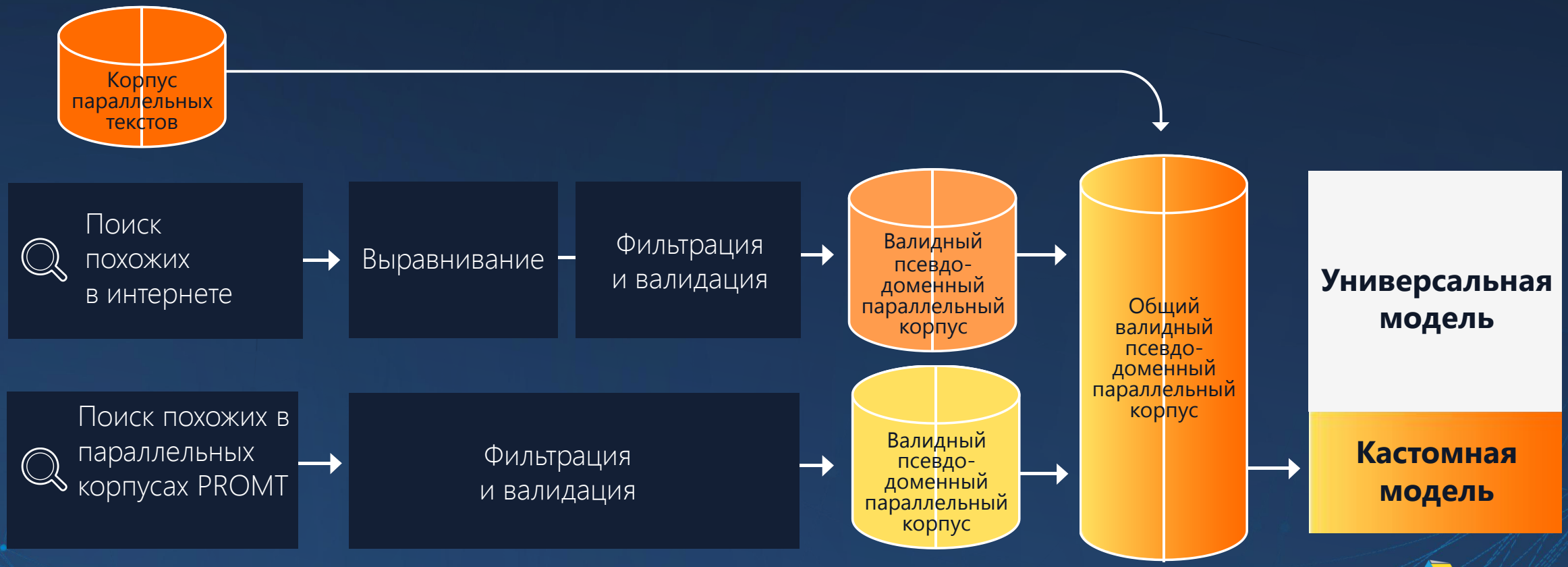
Заказчик дал двуязычные корпуса **достаточного** объема



PROMT

Тренировка PROMT Neural

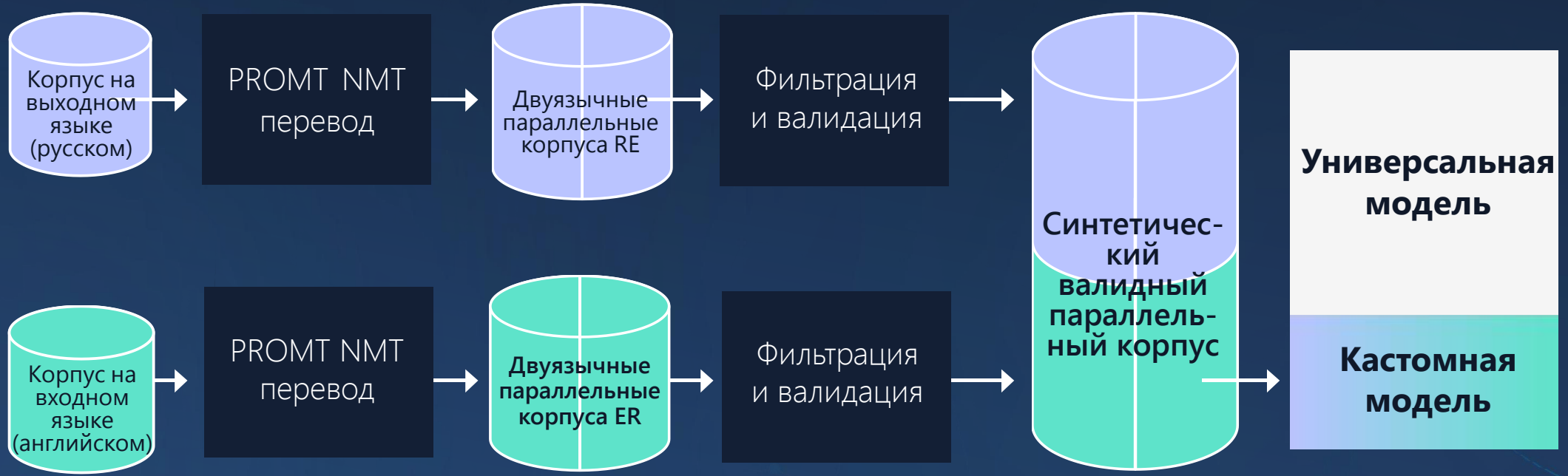
Заказчик дал двуязычные корпуса **незначительного** объема



PROMT

Тренировка PROMT Neural

Заказчик **не дал** параллельных корпусов



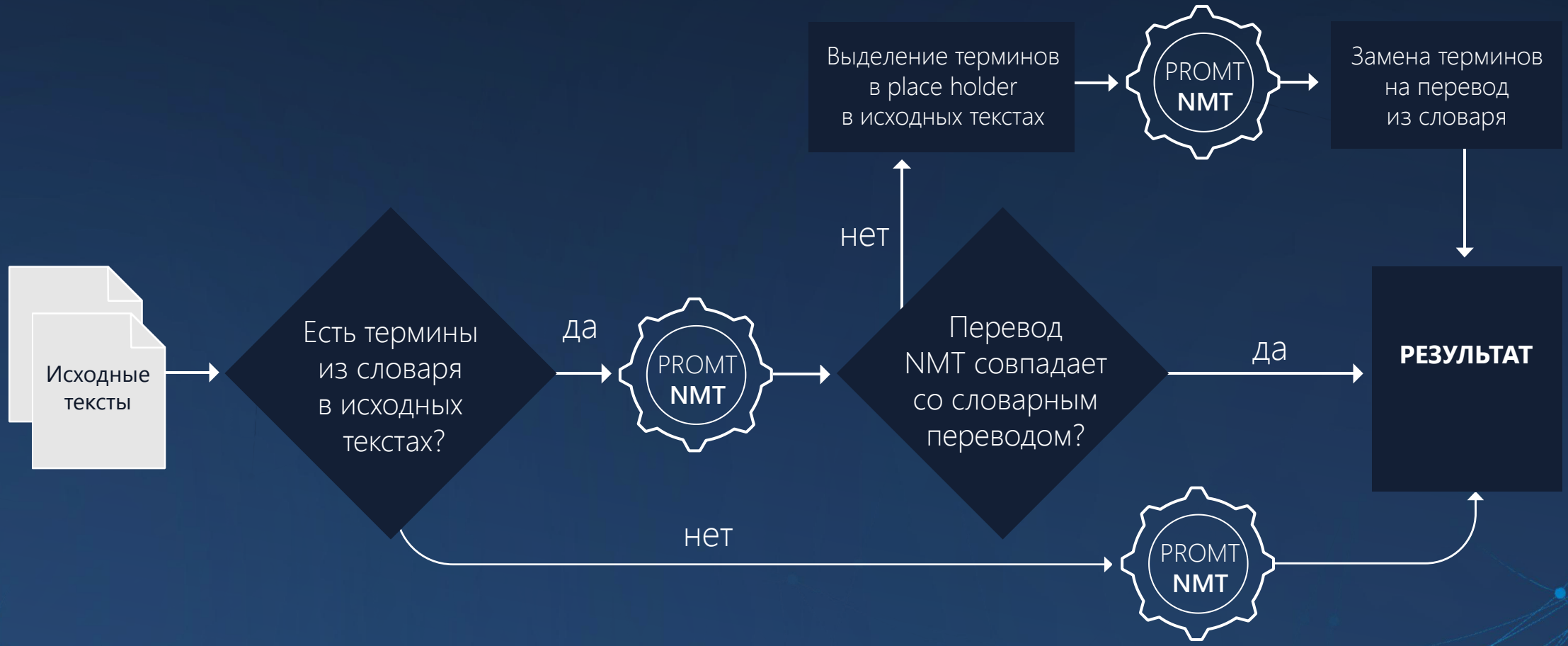
PROMT

Тренировка PROMT Neural

Заказчик предоставил **глоссарии**

PROMT

Схема перевода со словарем



PROMT

ПРОДУКТ

PROMT Neural Translation Server

PROMT Neural Translation Server

Параметры

- Свой сервер/Облако
- Windows/Linux

Рекомендуемые аппаратные требования:

- процессор с 4 и более ядрами;
- 16ГБ RAM;
- CUDA-совместимая видеокарта с 4GB видеопамяти (рекомендуется класса Nvidia GeForce 1080 Ti или выше)
- 100ГБ свободного места на диске

PROMT

Результат перевода Атомная энергия

Корпуса для настройки:

До валидации -
530 000 сегментов

После валидации –
274 972 сегментов

Bleu Scores

До настройки - 35.18
После настройки - 50.98

Post-editing distance

Универсальная модель - 50,2%
Кастомная модель - 37%

Результат перевода Атомная энергия

Оригинал:

The hydraulic lock located on the side of the fuel pool is designed to maintain a water level at elevation + 25.450 m in the fuel pool during all modes of NPP operation

До настройки (PROMT NMT):

Гидравлический замок, расположенный со стороны топливного бассейна, предназначен для поддержания уровня воды на высоте + 25,450 м в топливном бассейне на всех режимах работы АЭС.

Эталон:

Гидрозатвор, расположенный со стороны самого бассейна выдержки, предназначен для поддержания уровня воды на отметке + 25,450 м в бассейне выдержки во всех режимах эксплуатации АЭС.

После настройки (PROMT NMT) :

Гидрозатвор, расположенный со стороны бассейна выдержки, предназначен для поддержания уровня воды на отметке + 25.450 м в бассейне выдержки во всех режимах работы АЭС.

Тренды и вызовы в NMT

- Работа с терминологией
- Перевод документов
- Новые метрики для оценки качества перевода
- Развертывание решений

Тренды и вызовы для индустрии

- Сбор и подготовка данных
- Контроль качества перевода
- Больше постредактирования, чем перевода

Ваши вопросы?

Написать и получить демо-доступ

Юлия Епифанцева
julia.epiphantseva@promt.ru