

Нейронный машинный перевод:

одно универсальное
решение или много
специализированных?

Опыт
PROMT Neural



Краткая история машинного перевода

Rule-Based MT

Машинный акцент в переводе, предсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ
СЛОВАРИ

С 1950 гг.

Statistical MT

Более гладкий перевод, чем в RBMT, но непредсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ
ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ

С 2000 гг.

Neural MT

Абсолютно гладкий перевод, иногда непредсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ
ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ

С 2015 гг.

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

Neural MT для индустрии перевода - это

- ТЕХНОЛОГИЯ
- ДАННЫЕ
- ПРОДУКТ

The background of the slide is a complex, abstract digital circuit. It features a dense network of glowing blue lines, nodes, and symbols on a dark blue gradient. The elements include various geometric shapes like circles, squares, and triangles, as well as symbols such as plus signs, arrows, and binary-like patterns. Some lines are thicker and more prominent, while others are thin and delicate. The overall effect is a sense of high-tech connectivity and data flow.

ТЕХНОЛОГИЯ

PROMT Neural

Гибридная технология PROMT Neural



Трансформер,
Marian NMT



Алгоритмы глубокого
семантико-синтаксического
анализа

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

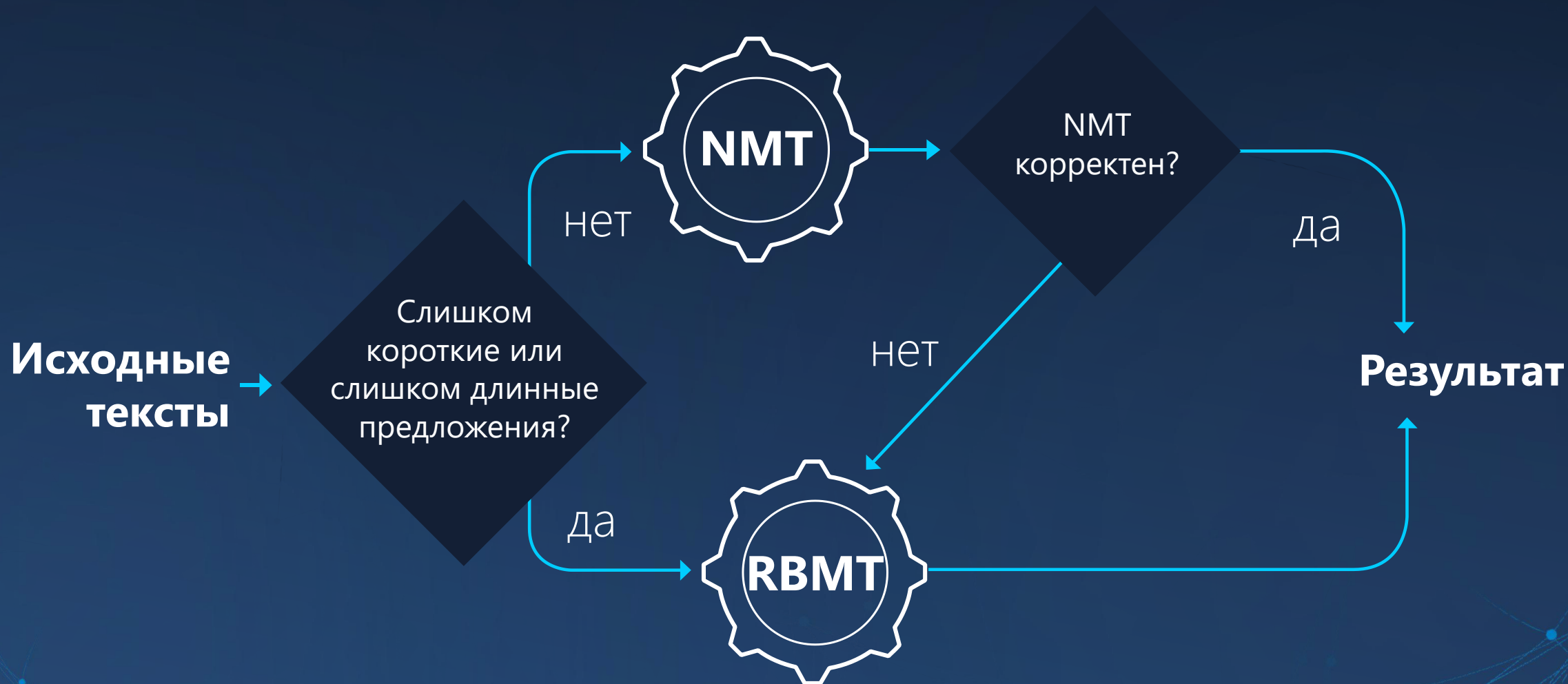
«РАЗМЕЧЕННЫЕ»

«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

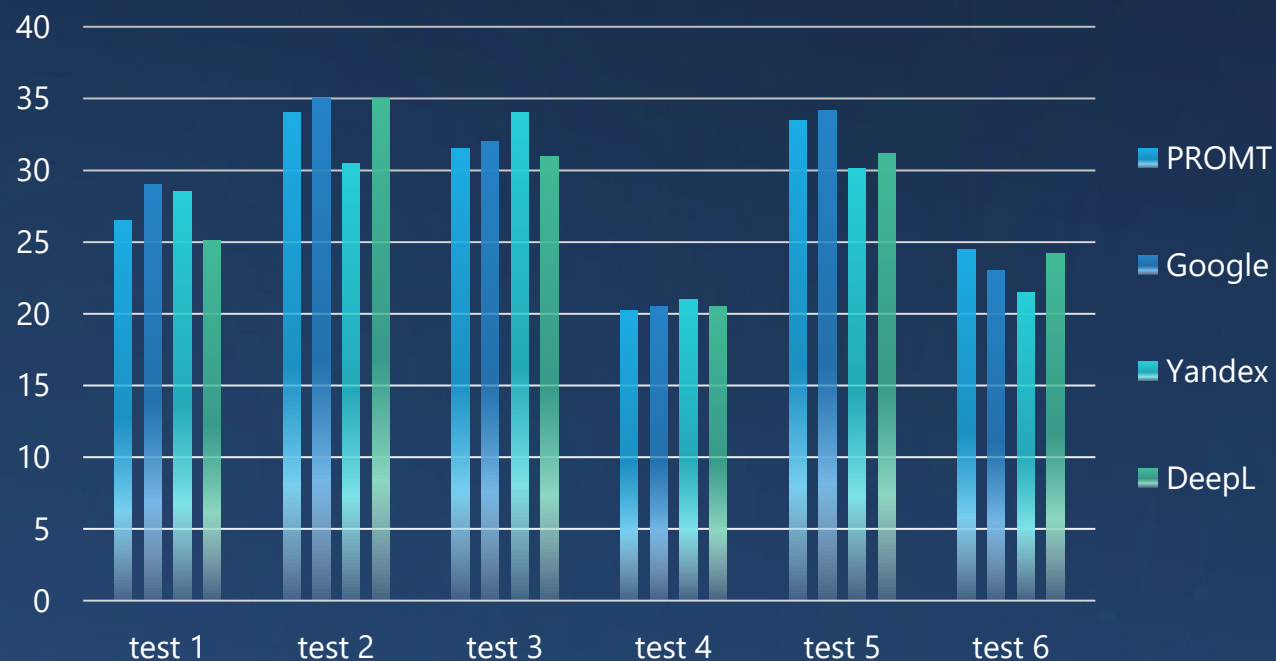
Как работает PROMT Neural?



PROMT Neural, Google, Yandex, DeepL

Тестирование на текстах общей тематики, ER

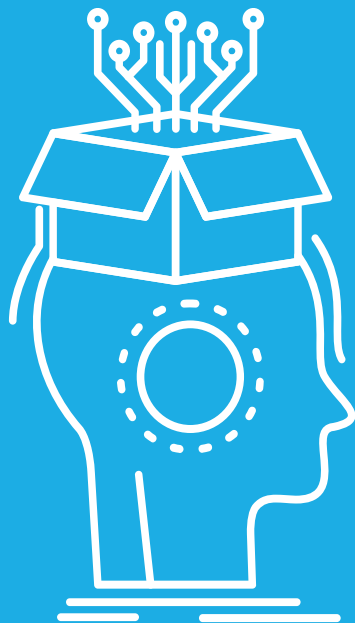
BLEU Scores



Экспертная оценка

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

Задача перевода решена?



- **Специализированные тексты** переводятся не очень хорошо
- **Обучение**
На чем тренировать NMT?
Сколько нужно данных? Что делать, если нет параллельных данных? Можно ли тренировать на глоссариях?

The background of the slide is a complex, abstract digital circuit. It features a dense network of glowing blue lines, nodes, and geometric shapes like circles, squares, and triangles, all set against a dark blue gradient. The lines and nodes are interconnected, creating a sense of data flow and connectivity. Some elements are highlighted with brighter blue or white light, giving the impression of active components or data paths.

ДАННЫЕ

ДВУЯЗЫЧНЫЕ И МОНОЯЗЫЧНЫЕ
КОРПУСА И ГЛОССАРИИ
ЗАКАЗЧИКА

Типичные ошибки NMT

- Перевод имен собственных
- Нарушение единства терминологии
- Нейрологизмы

Почти все проблемы решаются тренировкой на релевантных по объему и тематике данных

Что такое тематически однородные данные?

Тематика: нефть/газ, энергетика

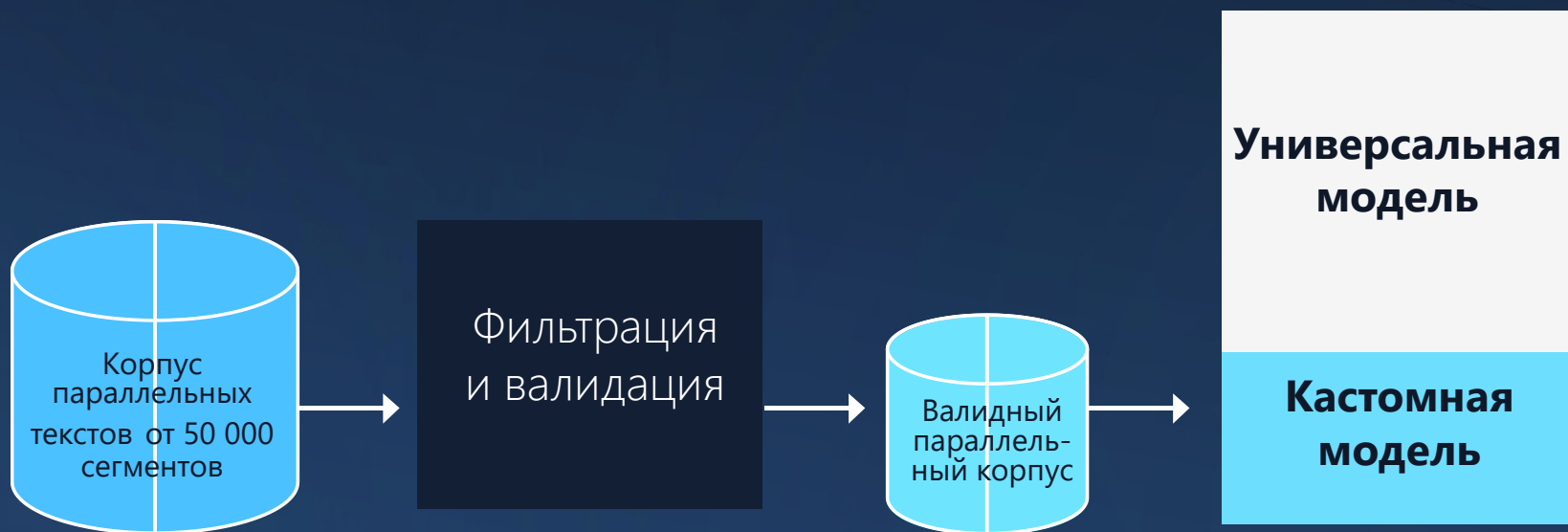
Тип текста: договоры, инструкции, User-generated Content...

Заказчик особенности текстов разных компаний

Переводчик: особенности индивидуального стиля

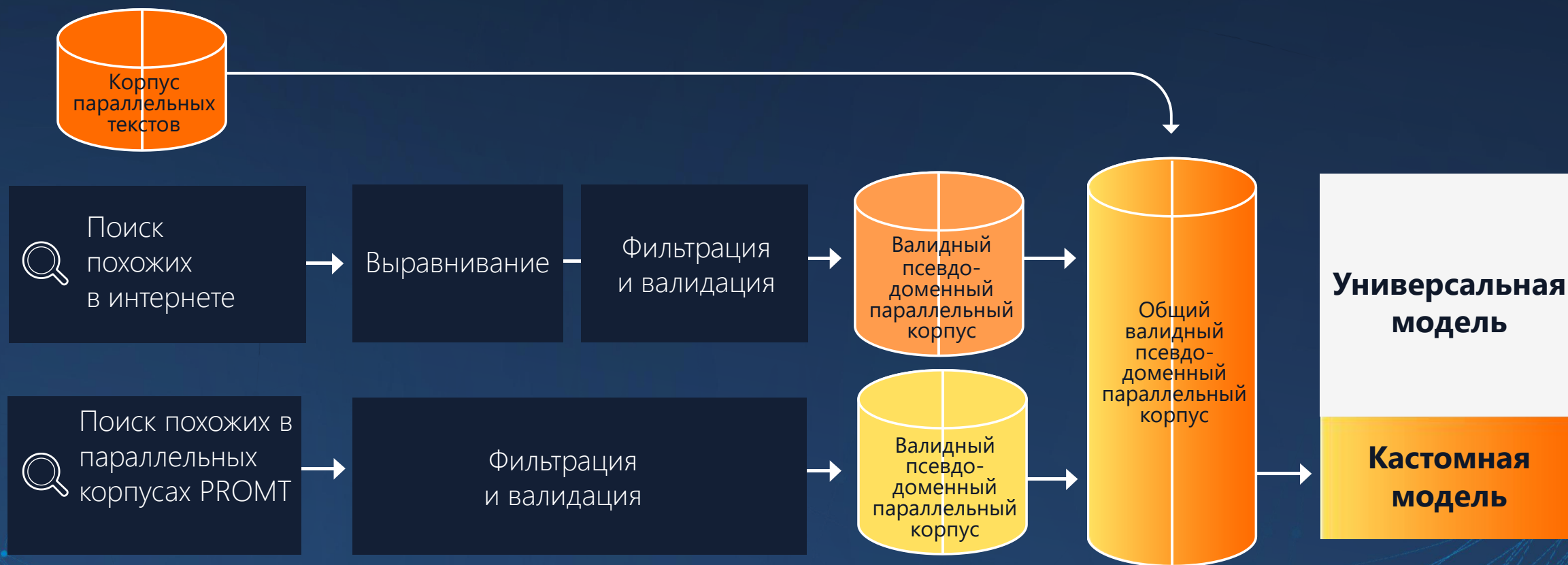
Тренировка PROMT Neural

Заказчик дал двуязычные корпуса **достаточного** объема



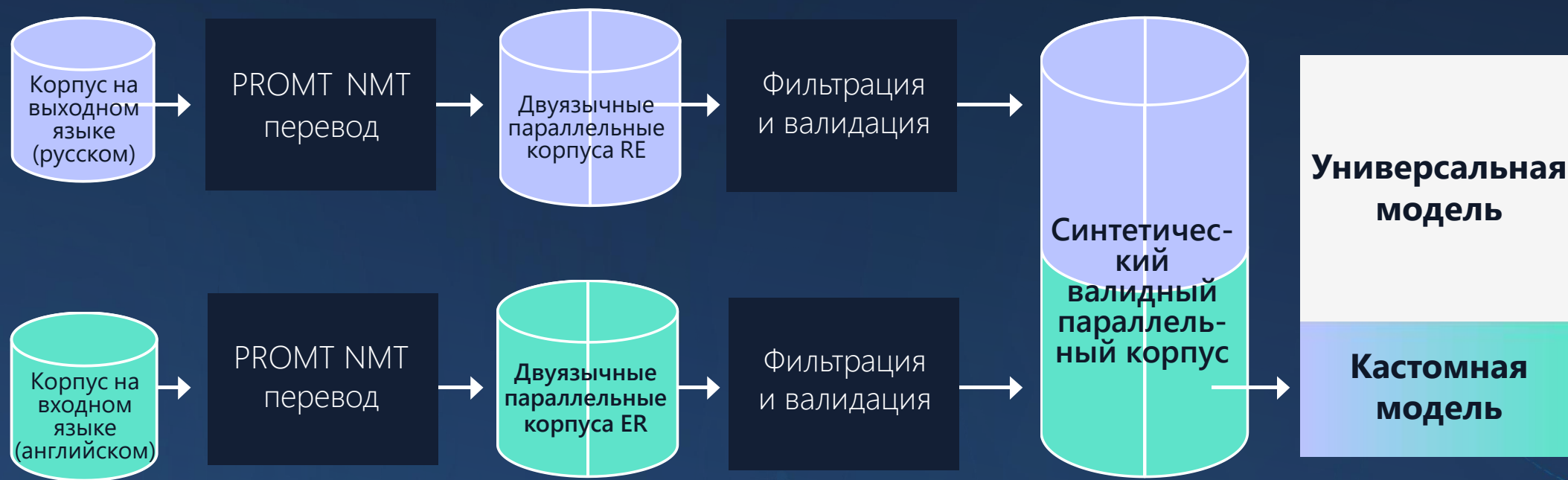
Тренировка PROMT Neural

Заказчик дал двуязычные корпуса **незначительного** объема



Тренировка PROMT Neural

Заказчик **не дал** параллельных корпусов



Тренировка PROMT Neural

Заказчик предоставил **гlossарии**



Схема перевода со словарем



The background of the slide is a complex, abstract digital circuit. It features a dense network of glowing blue lines, nodes, and symbols on a dark blue background. The elements include various geometric shapes like circles, squares, and triangles, as well as symbols like plus signs, arrows, and binary digits (0s and 1s). The overall effect is a sense of high-tech connectivity and data flow.

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

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@prompt.ru

