СИНЕРГИЯ ДАННЫХ И ТЕХНОЛОГИЙ ДЛЯ МАШИННОГО ПЕРЕВОДА И ОБРАБОТКИ ЯЗЫКА

EGOTECH PROMT

Спикеры:
Юлия Епифанцева, PROMT
Маргарита Меняйлова, ГК ЭГО Транслейтинг

О чем мы говорим

- Возможности применения данных безграничны
- Возможности MT шире стандартного понимания
- Данные и MT как слаженный механизм обеспечения КАЧЕСТВА и КОНФИДЕЦИАЛЬНОСТИ
- Инструменты работы с данными от EGOTECH (Компания ЭГО Транслейтинг)
- Кастомизированный PROMT NMT
- Эксперимент: симбиоз технологий для роста качества NMT

EGOTECH PROMT

PROMT и ГК ЭГО Транслейтинг смотрят в будущее

Почему PROMT и EGO Translating поднимают тему синергии данных и МТ?

- Наши компании на практике осознали необходимость совмещать продуманную работу с данными и грамотную настройку МТ
- Взаимная заинтересованность в синергии технологий и специалистов
- У нас есть техническая возможность реализации подобного совмещения!
- Качество совместной работы говорит само за себя

EGOTECH PROMT

ДАННЫЕ

ВОЗМОЖНОСТИ И ВЫЗОВЫ

Ключевые вопросы данных

ДАННЫЕ - электричество / нефть 21 века.

90% данных сгенерировано за последние 2 года

80% данных во всех отраслях – **текст**

2025 г – 80% всех данных – неструктурированы (ВШЭ)

Ключевой вопрос данных – качество:

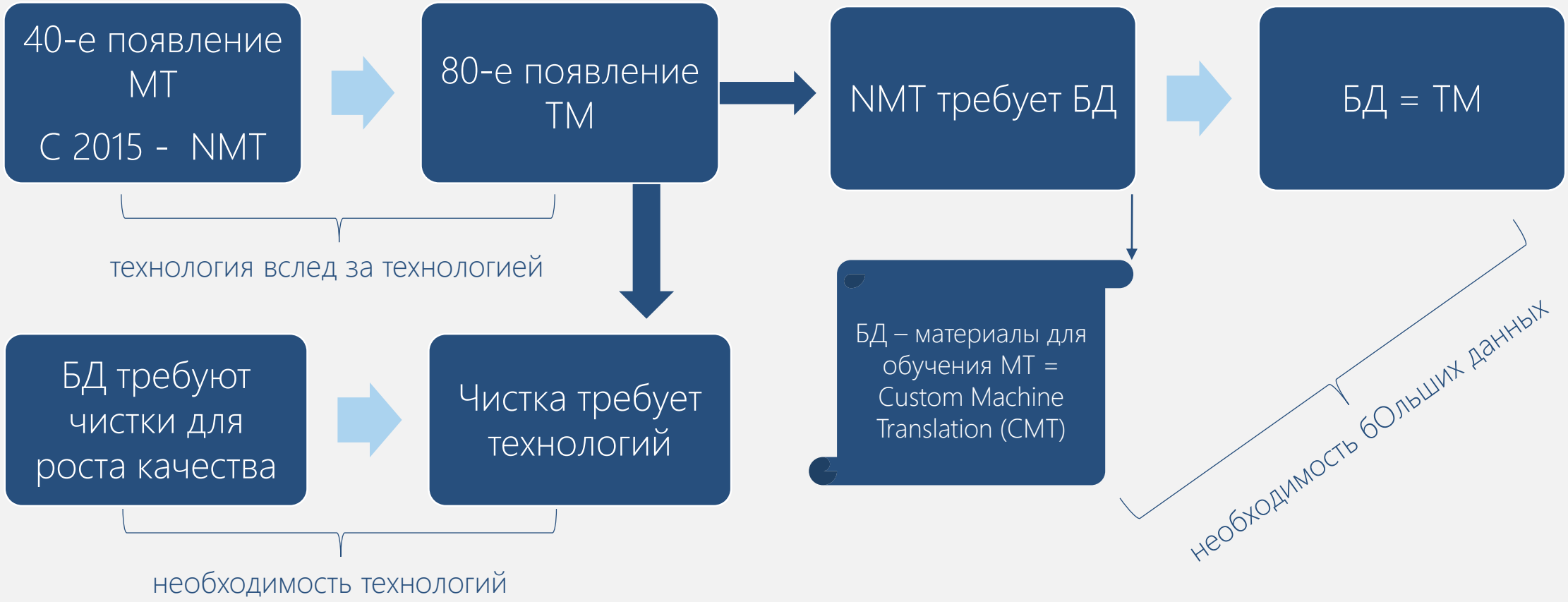
- Полнота
- Правильность
- Единообразие
- Обогащенность (данные + метаданные становятся информацией)

А также:

- Сохранность
- Конфиденциальность

EGOTECH PROMT

Больше качественных данных – лучше MT!



EGOTECH PROMT

ТРЕНДЫ МТ СЕГОДНЯ

MT сегодня

- Нейросетевые технологии: рекуррентные НС > трансформеры.
- 100+ разработчиков технологии NMT на основе opensource и собственных разработок, всего 2 - в РФ.
- Повышение роли данных и их качества для обучения нейросети. Проблема создания систем МТ для региональных и редких языков из-за нехватки данных.
- Проблема улучшения качества перевода существующих систем NMT как конкурентное преимущество.
- Проблема конфиденциальности. Локальный контур (в РФ только один поставщик решений МТ для работы в локальной сети).
- Возросший интерес к технологии и сопутствующим услугам в связи с пандемией, как следствие, вопросы применения МТ и БД стали острее.

Применение MT и COVID-кризис

- Рост роли МП как единственной возможности быстро донести информацию на региональном языке.

Государственные организации в Юго-восточной Азии (e.g. Сингапур). Используется MT+MTPE для донесения информации до каждого на его локальном языке в режиме онлайн.

- Перевод в e-commerce, способствующий повышению конверсии.

ASIA Digital (Сингапур) – крупная компания, предлагающая встраиваться в интернет-магазины с использованием MT и MT+MTPE для повышения конверсии. Считают, что будущее в локализации ecom за МП.

- Облачные клиенты PROMT: медицина и фармацевтика – рост объемов переводов с начала пандемии

Сервис для врачей, где интегрированы технологии перевода PROMT для перевода статей и другой медицинской и фармакологической информации из зарубежных источников. В марте и апреле общее количество запросов на перевод у увеличилось в 160 раз, а объемы перевода – в 55 раз!

Плюсы MT+MTPE: быстрее и дешевле обычного перевода, подходит для обработки информации от различных видов коммуникаций, качество выше простого MT, сглаживаются острые места MT (культурные аспекты, чистота языка), гораздо шире область применения.

MT vs MT+MTPE: MT быстрее и дешевле MT+MTPE, но без MTPE область применения очень ограничена, не может быть использован при работе с контентом бренда.

EGOTECH PROMT

Кастомизация

- Что такое кастомный движок?
- Как он создается?
- Какие компании предлагаю кастомизацию?
- Кастомизация возможна только на качественных данных большого объема!
- Стоимость кастомизации?
- Плюсы?

Custom Machine Translation – это технология + данные

Новая мощная синергия, которая логично появилась в связи с потребностями меняющегося мира

EGOTECH PROMT

EGOTECH

ТЕХНОЛОГИИ ДЛЯ
ОБРАБОТКИ ДАННЫХ

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

EGOTECH PROMT

EGOTECH
для решения следующих задач:

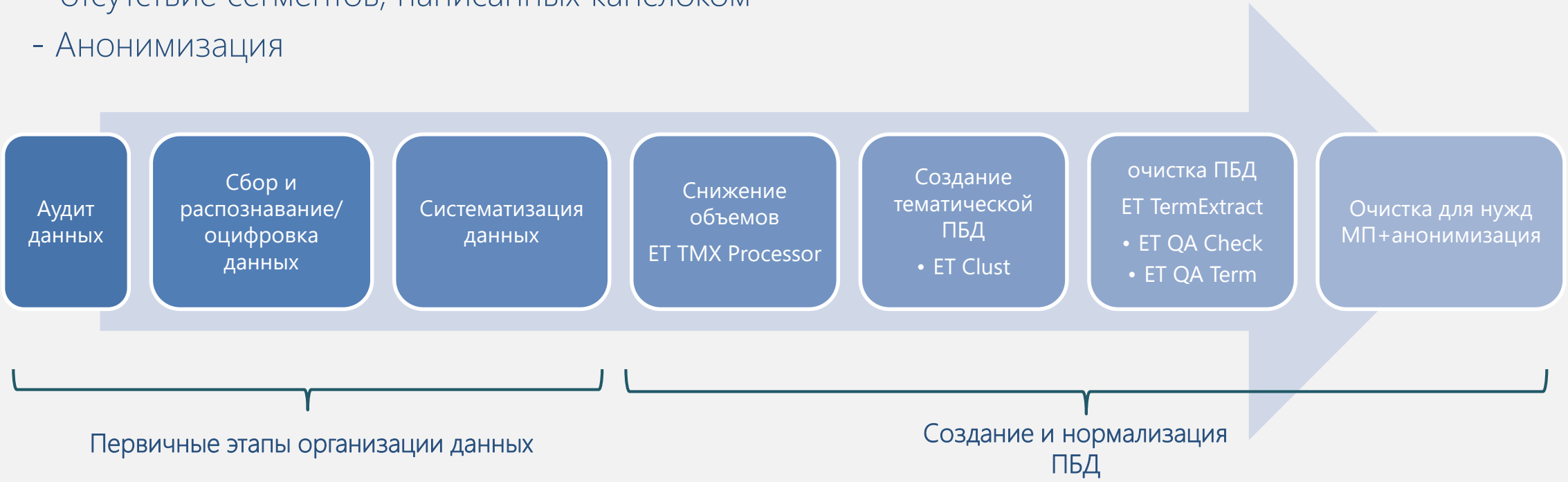
- Сбор, очистка и нормализация переводческих баз данных Translation Memory (далее – ПБД)
- Создание качественных одно- и многоязычных словарей
- Соблюдение правильности и единства терминологии в проектах
- Оптимизация работы со словарными материалами
- Оценка и обеспечение качества переводов
- Оптимизированное внедрение МП
- Адаптивная настройка системы МП с целью улучшения качества перевода

EGOTECH PROMT

Обработка БД для МП

Чем обучающий датасет отличается от переводческой Translation memory?

- Большие объемы (от 50 000 сегментов)
- Технические особенности: 1 сегмент на вход = 1 сегмент на выход; отсутствие сегментов, написанных капслоком
- Анонимизация



Результат

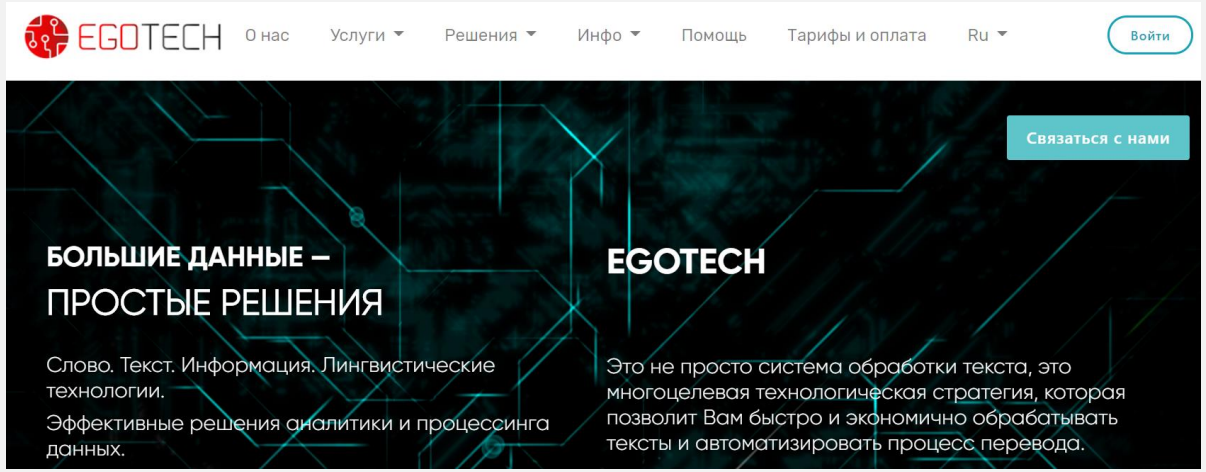
- ✓ Тематические БД
- ✓ БД, совместимые с основными CAT
- ✓ Облегченные БД
- ✓ БД с правильной терминологией
- ✓ Корректные БД с правильной нелингвистической информацией

- ✓ Снижение себестоимости и временных затрат на перевод до 70%
- ✓ Повышение качества выполнения работ до 50%

EGOTECH PROMT

Ключевые преимущества

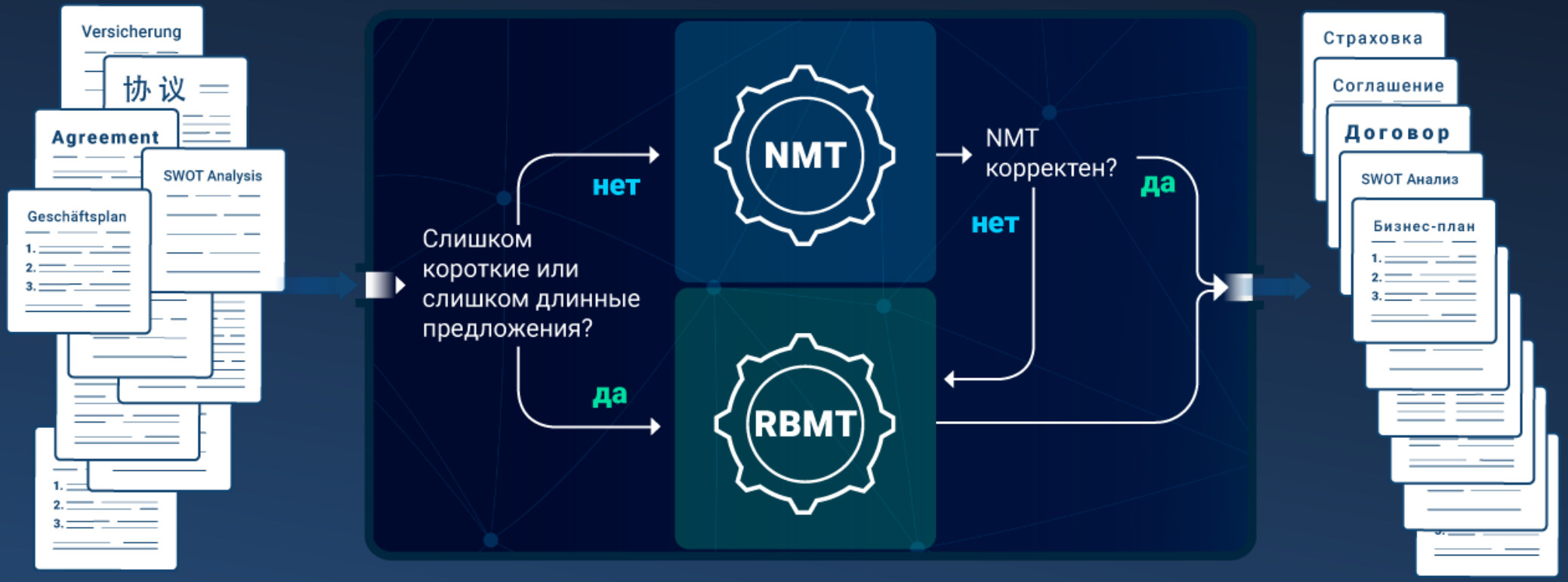
- ✓ Уникальность собственных разработок на основе глубокого анализа языков
- ✓ Решения, созданные на базе реальных осознанных потребностей
- ✓ Огромный опыт работы с переводческими процессами и данными
- ✓ Комплексность предлагаемых решений и услуг на их базе
- ✓ Платформенные и серверные версии
- ✓ Кастомизация под разные задачи
- ✓ Оптимальная стоимость



PROMT NEURAL

ТЕХНОЛОГИЯ NMT PROMT

Как работает PROMT Neural?



EGOTECH PROMT

Объемы данных PROMT NMT Универсальная модель, ER

Корпуса параллельных текстов

«ЧЕЛОВЕЧЕСКИЕ»

WMT, OPUS, PROMT корпуса
/ 50+ млн сегментов

«СИНТЕТИЧЕСКИЕ»

Wikipedia / 22 млн сегментов
Новости / 22 млн сегментов

«РАЗМЕЧЕННЫЕ»

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

EGOTECH PROMT

Почему универсальную модель нужно дополнительно обучать?

- Термины не встречаются или встречаются редко в общих данных
- Термины встречаются часто, но имеют «неправильный» (универсальный) перевод
- Термины встречаются часто, но имеют как «неправильный» (универсальный), так и правильный перевод

Универсальная модель всегда будет отдавать предпочтение универсальным (общим) переводам и не сможет корректно переводить специализированные термины

Методы обучения PROMT NMT

Дополнительное обучение NMT модели на обучающей выборке

Коррекция перевода отдельных терминов через глоссарий (словарь)

EGOTECH PROMT

Обучение

PROMT NMT модели на параллельных данных (TM)

Обучение PROMT NMT на параллельных данных. Требования

- Дополнительное обучение только на новой обучающей выборке
- Тематическая однородность данных
- Источник (корпоративные translation memories или тексты и их переводы с последующим выравниванием)
- Достаточный для обучения объем - от 100 000 сегментов

EGOTECH PROMT

Пример зависимости качества перевода от объема обучающей выборки

Оригинал:

The GCF system equipment, valves and mainstream piping refer to 4N safety class according to NP-001-15 and seismic category II as per NP-031-01.

Настройка на 10 000 предложениях:

Оборудование, клапаны и трубопроводы системы ГСФ относятся к классу безопасности | 4Н по НП-001-15 и сейсмической категории II по НП-031-01.

Универсальная модель PROMT NMT:

Оборудование, клапаны и трубопроводы системы РГС относятся к классу безопасности 4Н по НП-001-15 и сейсмической категории II по НП-031-01.

Настройка на 100 000 предложениях:

Оборудование, арматура и трубопроводы системы ГКУ относятся к классу безопасности 4Н по НП-001-15 и сейсмической категории II по НП-031-01.

Эталон:

Оборудование, арматура и трубопроводы основного потока системы GCF относятся к 4Н классу безопасности в соответствии с НП-001-15 и II категории сейсмостойкости по НП-031-01.

Настройка на 250 000 предложениях:

Оборудование, арматура и трубопроводы основного потока системы GCF относятся к классу безопасности 4Н по НП-001-15 и II категории сейсмостойкости по НП-031-01.

EGOTECH PROMT

Коррекция

переводов терминов через глоссарии

Коррекция переводов терминов в PROMT NMT

- Используется приложения PROMT с UI, специальное обучение не нужно
- 1 ключ - 1 перевод
- Поддержка слов и словосочетаний
- Только существительные, морфология определяется автоматически
- Ручной и автоматический режимы (ввод из файла)

EGOTECH PROMT

Пример настройки

Оригинал:

Why is Compliance Management so crucial to your business?

Глоссарий

Compliance Management – контроль за соблюдением

Перевод без глоссария:

Почему управление соответствием нормативным требованиям имеет столь важное значение для вашего бизнеса?

Перевод с глоссарием:

Почему контроль за соблюдением имеет столь важное значение для вашего бизнеса?

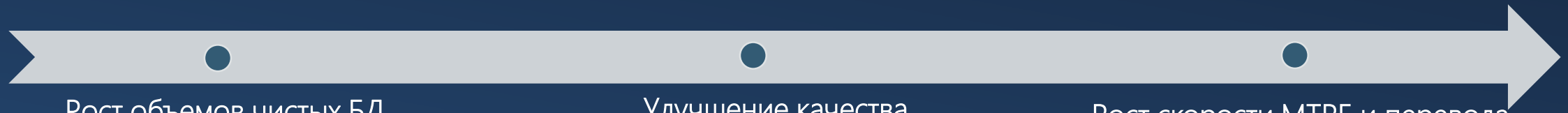
EGOTECH PROMT

Влияние объема данных на качество PROMT NMT

Количество выровненных предложений в новой обучающей выборке	Средний прирост BLEU scores
10 000 пар предложений	+0,3
50 000 пар предложений	+3
100 000 пар предложений	+6
250 000 пар предложений	+10

Анализ кастомной модели с очищенными БД

БД 10 000 сегментов	Незначительное улучшение качества МП	Нормативы постредактирования: 20 стр/день
БД 50 000 – 100 000 сегментов	Существенное улучшение качества МП	Нормативы постредактирования: 25 стр/день
БД от 500 000 сегментов	качество МП, сопоставимое с качеством обычного перевода, но с увеличением скорости обработки в 5 раз	Нормативы постредактирования: 50 стр/день



Рост объемов чистых БД

Улучшение качества итогового перевода

Рост скорости MTPE и перевода в целом

* Расчет производился Компанией ЭГО Транслейтинг на основе анализа скорости MTPE на разных языковых парах (Google)

EGOTECH PROMT

PROMT Neural Translation Server

Параметры

- Свой сервер/Облако
- Windows/Linux

Рекомендуемые аппаратные требования:

- 8 и более ядер (Intel Core i7 или выше);
- 32 ГБ RAM;
- CUDA-совместимая видеокарта с 4GB видеопамяти
- 100ГБ свободного места на диске

EGOTECH PROMT

PROMT Professional Neural

Параметры

- Десктоп Windows 7 SP1 (64x)

Рекомендуемые аппаратные требования:

- процессор Intel Core i5-44xx и старше с 6 и более ядрами;
- 16ГБ RAM;
- * CUDA-совместимая видеокарта с 4GB видеопамяти (рекомендуется класса Nvidia GeForce GTX1050)
- 10ГБ свободного места на диске

EGOTECH PROMT

Ключевые преимущества

- ✓ Серверные и десктопные PROMT NMT решения
- ✓ Впечатляющее стартовое качество
- ✓ Настройка «под ключ» на стороне PROMT или настройка на стороне клиента (базы TM, глоссарии)
- ✓ Интеграция с популярными CAT-Системами - SDL Trados, Memsoure, Across (сервер) и SDL Trados (десктоп)
- ✓ Работа в локальной сети, интернет не требуется

EGOTECH PROMT

ЭКСПЕРИМЕНТЫ

НАСТРОЙКА НА РАЗНЫХ ТИПАХ ДАННЫХ

РЕЗУЛЬТАТЫ НАСТРОЙКИ

Прочистка данных

- корректная терминология,
- единообразный перевод терминологии по всему тексту,
- отсутствие орфографических, пунктуационных ошибок,
- отсутствие "странных" сегментов (с тегами, словами верблюдами, буквами в неправильной раскладке, например, k в русском сегменте вместо к)

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

TM для настройки:

Непрочищенные -
76 164 сегментов

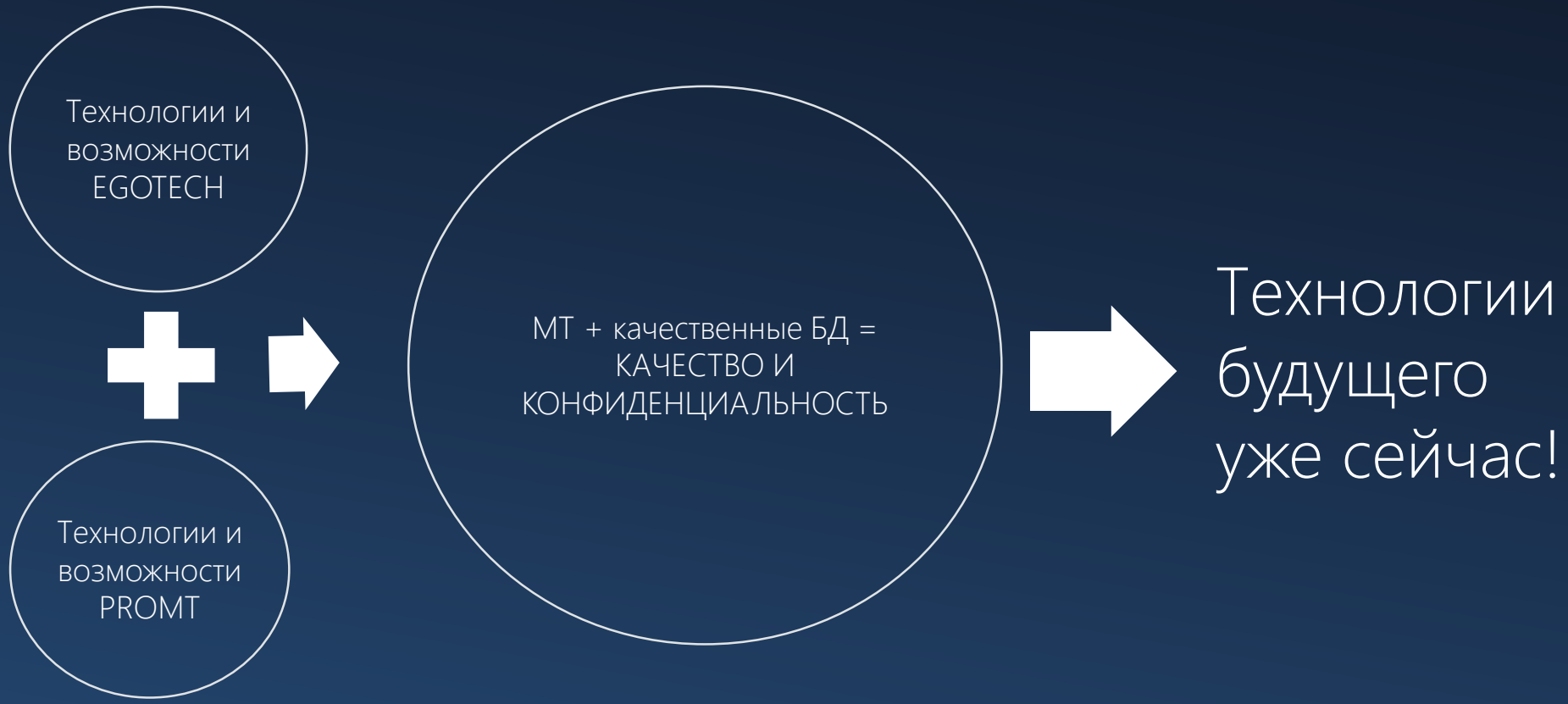
Прочищенные –
57 754 сегмента

Bleu Scores

До настройки - **44**.94
После настройки - 50.77

EGOTECH PROMT

ИТОГ



EGOTECH PROMT

Ваши вопросы?

https://egotech.tech/
Julia.Epiphantseva@promt.ru