НЕЙРОСЕТЕВОЙ МАШИННЫЙ ПЕРЕВОД

Как обучить модель МП
На примере PROMT Neural

Юлия Епифанцева
Директор по развитию PROMT

NMT

ОЧЕНЬ КРАТКО О ТЕХНОЛОГИИ

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 И ГЛОССАРИИ (СЛОВАРИ)

С 2015 гг.

PROMT

Что такое NMT?

Нейросетевой машинный перевод (Neural Machine Translation, NMT) – это автоматический процесс перевода текстов с одного естественного языка на другой с помощью математической модели, основанной на нейронных сетях.

Модель обучается на корпусах параллельных текстов (текстах на оригинальном языке и их переводах, выровненных по предложениям) очень большого объема (большего, чем в SMT)

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

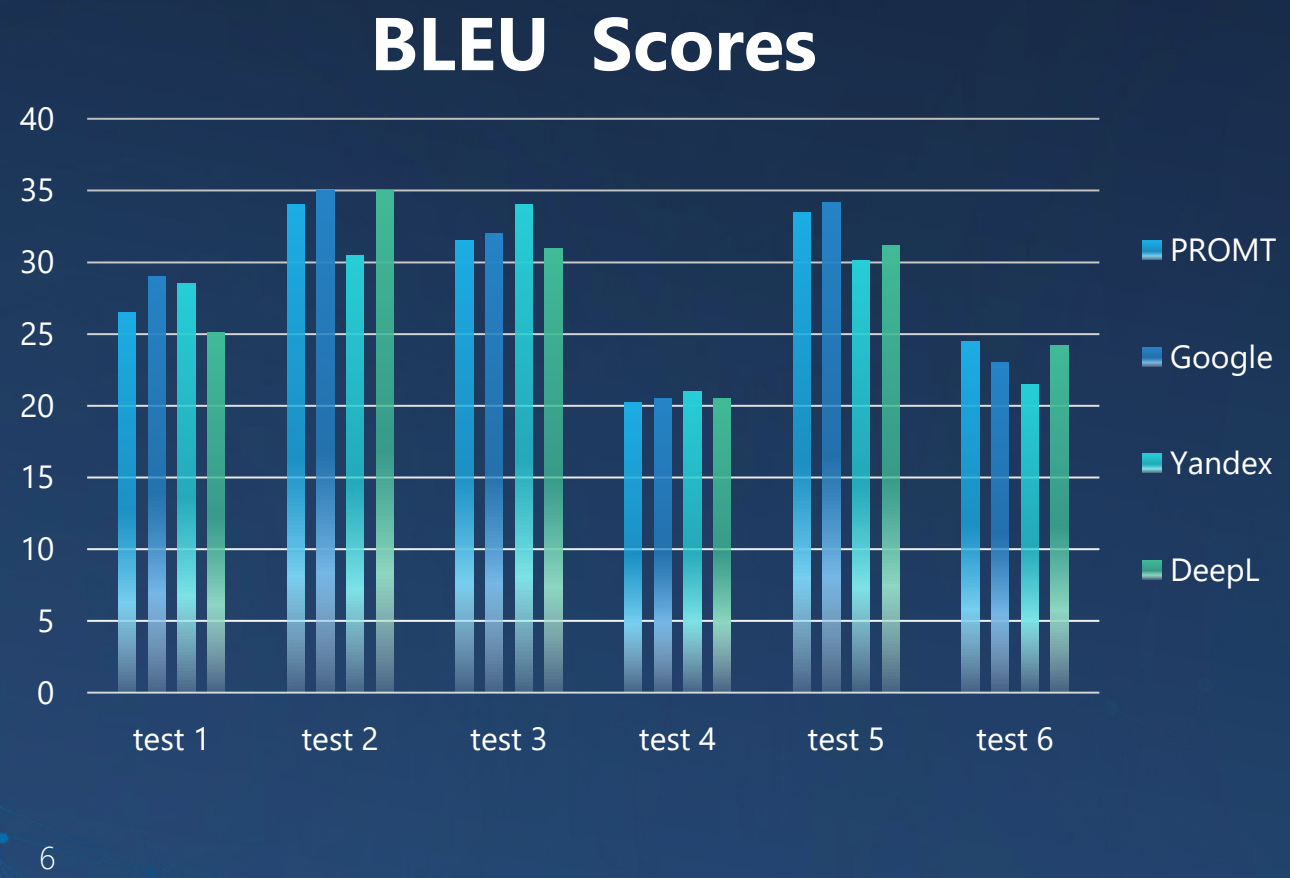
«РАЗМЕЧЕННЫЕ»

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

PROMT

PROMT Neural, Google, Yandex, DeepL

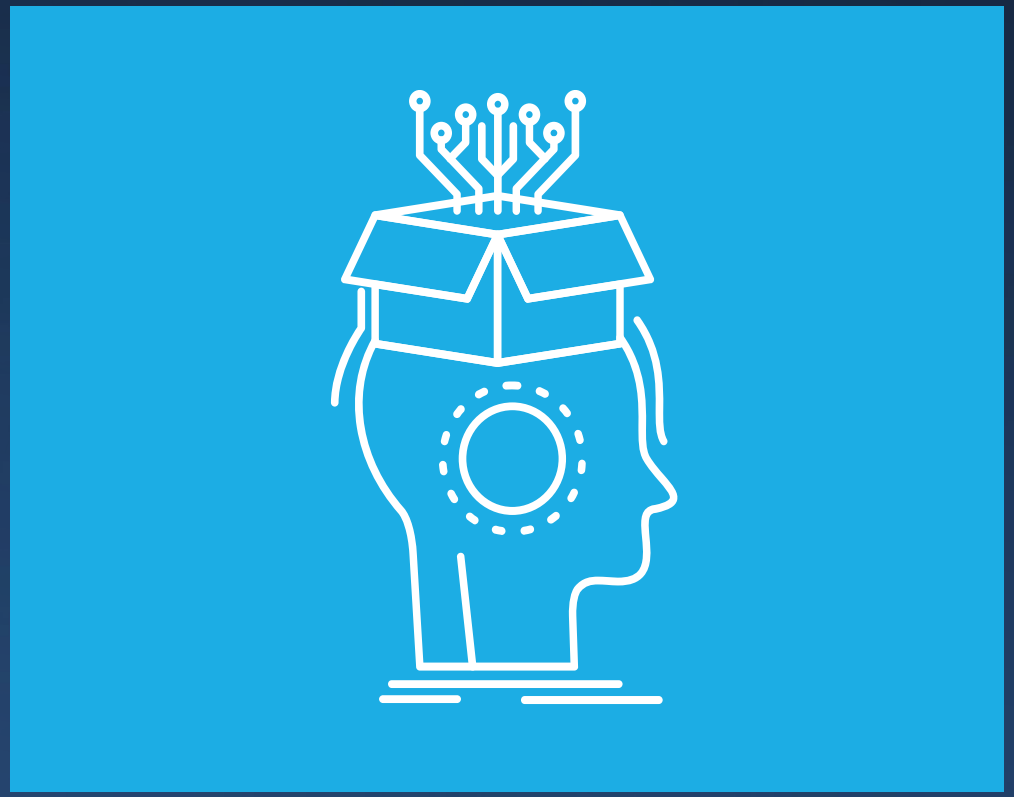
Тестирование на текстах общей тематики, ER



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
- **Настройка**
  На чем настраивать NMT? Сколько нужно данных? Что делать, если нет параллельных данных? Можно ли делать настройку на глоссариях?

PROMT

НАСТРОЙКА

ОБЩАЯ ИНФОРМАЦИЯ

Что такое настройка NMT?

Настройка NMT модели - это дополнительное обучение модели для повышения качества перевода текстов, близких по тематике, стилю и терминологии к данным, на которых модель будет обучаться

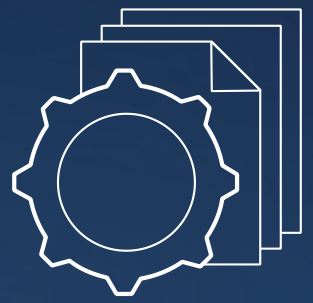
Почему универсальную модель нужно обучать?

- 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

PROMT

Методы обучения PROMT NMT



Дополнительное обучение NMT модели на обучающей выборке

Коррекция перевода отдельных терминов через глоссарий (словарь)

PROMT

НАСТРОЙКА

ОБУЧЕНИЕ PROMT NMT МОДЕЛИ НА ПАРАЛЛЕЛЬНЫХ ДАННЫХ

Обучение PROMT NMT на параллельных данных. Требования

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

Что такое тематически однородные данные?

Тематика:
двигатели в автомобилях ≠ авиационные двигатели….

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

Заказчик особенности текстов разных компаний

Переводчик: особенности индивидуального стиля

Translation Memories

- Базы переводов клиентов, создаваемые в CAT инструментах (выровнены по предложениям)
- Формат tmx
- (!) Часть данных будет отбракована из-за «технических» ошибок

Тексты и их переводы

- Необходимо выравнивание по предложению (полуавтоматический процесс)
- Качество выравнивания зависит от сложности формата и типа текста
- (!) Останется только 50-70% от первоначального объема

Что такое достаточный объем? Эксперимент 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

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.

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 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.

Настройка на 50 000 предложениях :

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

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.

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 100 000 предложениях:

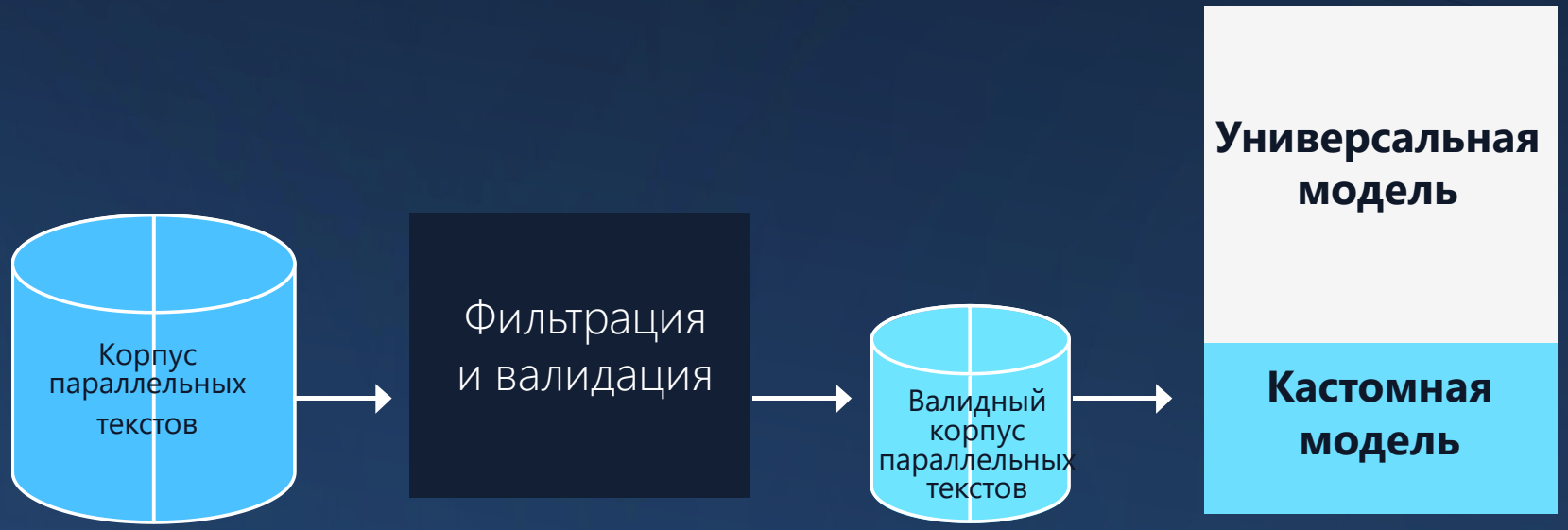
Оборудование, арматура и трубопроводы системы ГКУ относятся к классу безопасности 4Н по НП-001-15 и сейсмической категории II по НП-031-01.

Настройка на 250 000 предложениях :

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

Обучение PROMT NMT

на корпусе параллельных текстов



PROMT

Пример настройки

Данные для настройки:

До валидации -
530 000 предложений (tmx)

После валидации –
274 972 предложения

Bleu Scores

До настройки - 35.18
После настройки - 50.98

Post-editing distance

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

Пример настройки

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.

НАСТРОЙКА

КОРРЕКЦИЯ ПЕРЕВОДОВ ТЕРМИНОВ В PROMT NMT ЧЕРЕЗ ГЛОССАРИИ

Коррекция переводов терминов в PROMT NMT. Требования&Возможности

- 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)

Создание словаря в PROMT Neural

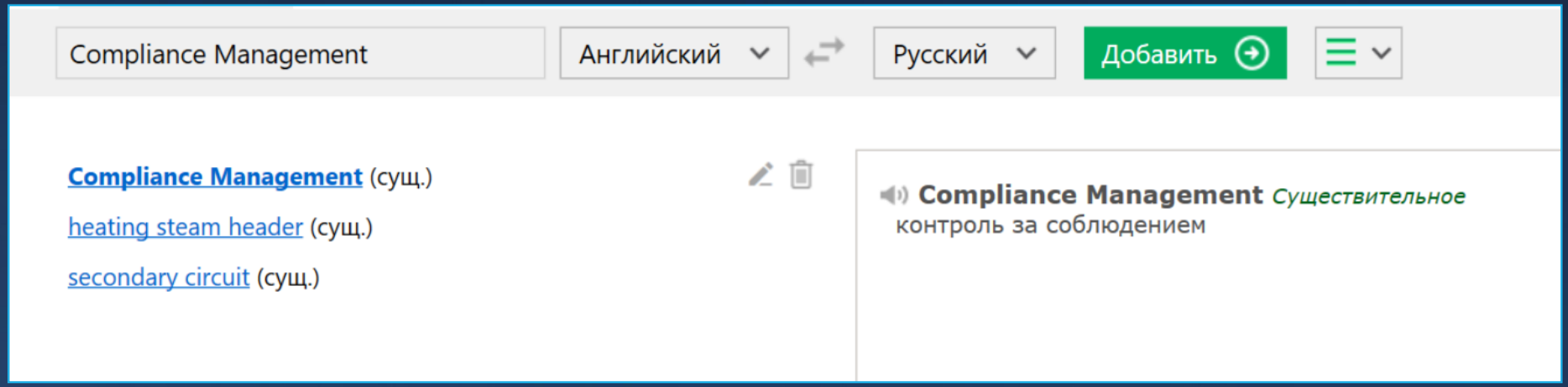
Из клиентских глоссариев

PROMT

Схема перевода со словарем

PROMT

Интерфейс PROMT для добавления терминов из глоссария



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?

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

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

Ваши вопросы?

Написать и получить демо-доступ

Юлия Епифанцева
julia.epiphantseva@promt.ru