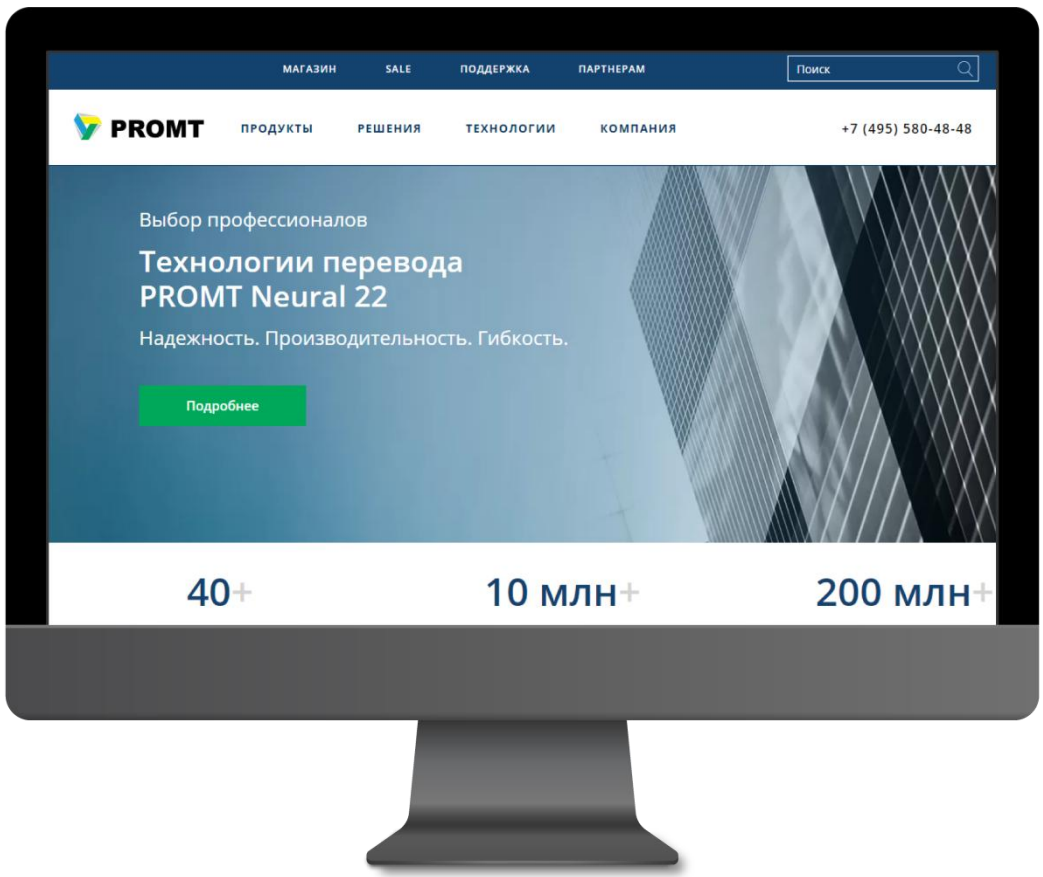
ТЕХНОЛОГИИ НЕЙРОНОГО МАШИННОГО ПЕРЕВОДА PROMT

Настройка системы машинного перевода без тренировочных данных заказчика

Юлия Епифанцева
PROMT

О компании

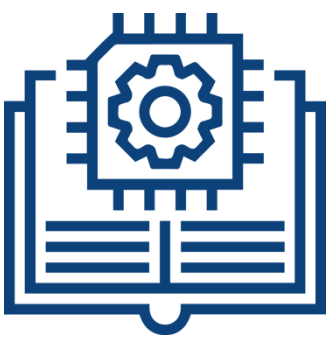
- Более 30 лет опыта в разработке и внедрении систем МП
- Один из крупнейших разработчиков ПО по версии РУССОФТ
- PROMT входит в топ-15 компаний в области обработки естественного языка и синтеза
- Постоянный участник Международной конференции WMT с 2013г.
- ИТ-решения для российских компаний



Методы обучения системы нейронного машинного перевода

Обучение на параллельных данных

Настройка системы машинного перевода на перевод терминологии и стилистических клише, уникальных для предметной области заказчика, на параллельных данных, т.е. ранее сделанных перевода.



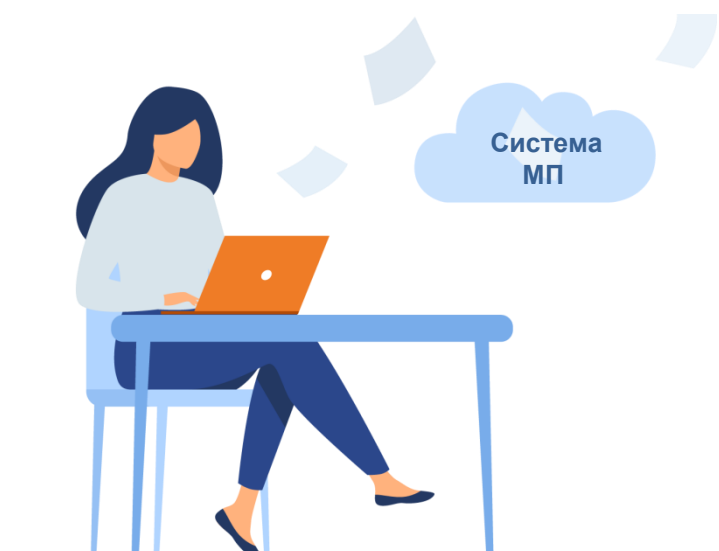
Настройка на глоссариях

Точный перевод узкоспециализированной терминологии благодаря интеграции терминологических глоссариев заказчика в систему машинного перевода.

Миф и реальность про обучение

МИФ

Любая система машинного перевода (например, онлайн-сервис, (само)обучается в то время, как пользователи переводят свои тексты.



РЕАЛЬНОСТЬ

Обучение системы - это отдельный процесс

Особенности обучения

- ✓ Подготовка данных в соответствии с требованиями
- ✓ Использование ПО для обучения
- ✓ Создание новой модели перевода и обновление системы МП

Сложности обучения на параллельных данных

Подготовка данных

- Высокие требования к объемам
- Принадлежность к одной предметной области
- Требования к качеству данных

  О фильтрации и валидации данных см. презентацию PROMT на TFR 2021
- Оценка результатов настройки (владение метриками)
- Квалификация сотрудников
- Время на сбор и/или обработку данных, автоматическую оценку результатов настройки

Обучение в офлайн -режиме

- PROMT Training Add-on для тренировки в офлайн режиме
- Высокие требования к аппаратным мощностям для развертывания ПО для обучения
- Стоимость ПО

Обучение в онлайн –режиме

- Риски утечки конфиденциальных данных на сторонних ресурсах

PROMT

ОБУЧЕНИЕ СИСТЕМЫ МП НА ГЛОССАРИЯХ

Что такое глоссарий машинного перевода?

- ✓ Список слов и словосочетаний с переводами
- ✓ Предназначен для добавления в систему МП
- ✓ Подготовлен по определенным правилам

Почему глоссарий важен для настройки?

Параллельные данные

Глоссарии

Как работает глоссарий в системах PROMT?

PROMT Neural Dictionary – специальная технология нейронного словаря, реализованная в нейронных продуктах системы перевода PROMT и доступная для Windows- и Linux-систем.

SMART Neural Dictionary

- учитывает контекст
- термины в переводе в нужном роде, числе и падеже

EN ↔ RU
EN ↔ FR
EN ↔ DE
ES → RU
FR → RU
PT → RU

SIMPLE Neural Dictionary

- без учета контекста
- термины остаются в той же форме, что и в глоссарии
- 10+ языков, 45+ языковых пар

Примеры работы глоссария в системах PROMT

SMART Neural Dictionary

Исходный текст	Перевод без глоссария	Перевод с глоссарием
The cable ladder is preferred to ensure cooling cables.	Кабельная лестница предпочтительна для обеспечения охлаждающих кабелей.	Предпочтение отдается лестничному лотку для обеспечения охлаждающих кабелей.

Примеры работы глоссария в системах PROMT

SIMPLE Neural Dictionary

Исходный текст	Перевод без глоссария	Перевод с глоссарием
L'encombrement et les points de raccordement sont identiques pour tous les types de transmissions.	Platzbedarf und Anschlusspunkte sind für alle Getriebetypen identisch.	Einbauverhältnisse und Anschlusspunkte sind für alle Getriebetypen identisch.
Tenir compte des consignes techniques.	Beachten Sie die technischen Anweisungen.	Beachten Sie den technische Vorgaben. *Правильно: … die technischen Vorgaben*

Общие советы при подготовке глоссария

- Не стремитесь сделать глоссарий как можно больше.
  - ✓ Чем больше глоссарий, чем сложнее следить за качеством представленной в нем информации,
  - ✓ Размер глоссария может влиять на скорость перевода.
- Оставляйте в глоссарии только те слова и словосочетания, с чем ваша система МП не справляется. Помните, что глоссарий МП – это не справочник для лингвиста (переводчика), а прежде всего это дополнительные данные для системы МП.
- Для каждого термина (слова или словосочетания) должен быть один перевод.
- Добавляйте в глоссарий термины, относящиеся к допустимым частям речи для данной системы МП (для PROMT - это только существительные, прилагательные).
- Не допускайте в глоссарии опечаток, служебных или форматных символов, комментариев.
- Следите за регистром: нарицательные имена должны быть представлены в глоссарии с маленькой буквы, имена собственные - в соответствии с правилами правописания.
- Если используемая система МП не умеет работать с морфологией, учтите это обстоятельство при составлении глоссария и при постредактировании.

Рекомендации PROMT при составлении глоссария

Не рекомендуется добавлять в глоссарий

- ✓ Частотные слова с общелексическими переводами
  accident – авария
- ✓ Частотные слова с узкоспециальным переводом
  vessel – сосуд (ср. судно)

Рекомендуется добавлять в глоссарий

- ✓ Имена собственные
- ✓ Аббревиатуры
- ✓ Словосочетания

ПРИМЕРЫ ВНЕДРЕНИЯ

Пример внедрения. СИБУР

СИСТЕМА МАШИННОГО ПЕРЕВОДА

PROMT Neural Translation Server

Кол-во лицензий: без ограничения

ДАННЫЕ ДЛЯ НАСТРОЙКИ

Англо-русский глоссарий - 23 900 терминов

Русско-английский глоссарий – 23 600 терминов

Пример внедрения. СИБУР

guide clip	направляющий крепежный элемент

Plastic knife guide clip to help with sharpening angles.	Пластиковый ножевой направляющий зажим для облегчения углов заточки.	Пластиковый ножевой направляющий крепежный элемент для облегчения углов заточки.

Результат

Более 23 000 терминов на английском и 23 000 на русском переводятся в соответствии с требованиями заказчика во всех текстах, документах, презентациях, сайтах, переводимых с помощью системы машинного перевода PROMT всеми сотрудниками компании.

Пример внедрения. Банк России

СИСТЕМА МАШИННОГО ПЕРЕВОДА

PROMT Neural Translation Server

Кол-во лицензий: 50 одновременных пользователей/5000 учетных записей

ДАННЫЕ ДЛЯ НАСТРОЙКИ

94 000 параллельных сегментов из памяти переводов

Англо-русский глоссарий -

1 973 терминов

Пример внедрения. Банк России

exposure at default	величина кредитного требования

Exposure at default is the total value a bank is exposed to when a loan defaults.	Риск убытков по умолчанию - это общая стоимость, которой подвержен банк при дефолте ссуды.	Величина кредитного требования - это общая стоимость, которой подвержен банк при дефолте ссуды.

Результат

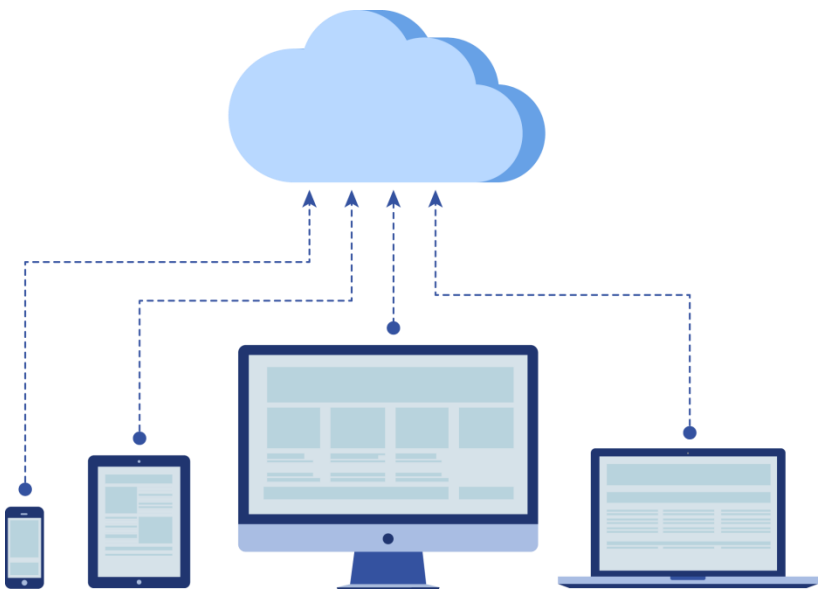
Рост качества перевода: для англо-русского перевода- с 30.41% на универсальной модели до 38.89% на тюнированной модели; - для русско-английского перевода - с 33.03% до 44.88%, перевод из глоссария присутствовал в 7,7% протестированных предложений.

Пользователи: общее количество уникальных пользователей: всего 1452, ежедневно более 100 пользуются корпоративной системой МП PROMT, а также сотрудники отдела переводов (ОП) Управления информационно-библиотечного обеспечения используют МП PROMT через CAT- систему.

Пример внедрения. TAdviser

СИСТЕМА МАШИННОГО ПЕРЕВОДА

PROMT Neural Translation Server по модели SaaS



ДАННЫЕ ДЛЯ НАСТРОЙКИ

Русско-английские глоссарии:

3 113 названий компаний

572 названия технологий/продуктов

2 722 имен персоналий

Пример внедрения. СИБУР

Ай-теко	I-Teco

«Ай-Теко» займется интеграцией ERP и АСУТП.	Ai-Teko will integrate ERP and PCS.	I-Teco will be engaged in integration ERP and PCS.

Результат

Портал TAdviser.com доступен любому пользователю интернета.

Аудитория портала TAdviser.com: по данным сервиса Similarweb.com на портале ежемесячно фиксируется от 55 000 до 113 000 визитов. География пользователей портала (топ-5 стран):США - 15%, Великобритания - 10%, Нидерланды – 9%, Россия - 5%, Вьетнам - 5%.

WMT 2021

Технологическая дорожка по настройке перевода по глоссарию

Results

EN-FR

	BLEU	EM	WO (2)	WO (3)	COMET	Rank (EM)	Rank (COMET)
Soft-constrained	47.69	**0.974**	**0.359**	**0.352**	0.752	**1-3**	3
SmartND	47.89	0.966	0.357	0.348	0.746	**1-3**	4-5
Best Scores	49.60	0.974	0.359	0.352	0.781	–	–

EN-RU

	BLEU	EM	WO (2)	WO (3)	COMET	Rank (EM)	Rank (COMET)
Soft-constrained	31.06	**0.909**	**0.254**	**0.255**	0.631	**1**	**1-2**
SmartND	**31.92**	0.788	0.243	0.241	**0.634**	10	**1-2**
Best Scores	31.92	0.909	0.254	0.255	0.634	–	–

- ✓ PROMT участвовала в настройке перевода с английского языка на французский и с английского на русский.
- ✓ В каждом направлении специалисты компании представляли результаты, полученные с помощью двух технологий настройки перевода терминологии: PROMT SmartND и исследовательской технологии "Перевод Терминологии с Ограничениями" (Soft-constrained).

Смотрим в будущее

Глоссарии MT являются доступным и эффективным способом повышения качества машинного перевода.

Преимущества глоссариев особенно заметны, если

- ✓ Нет параллельных данных для настройки
- ✓ Нет возможности передать данные для настройки третьим лицам или делать настройку inhouse,
- ✓ Настройка на параллельных данных не обеспечивает точный перевод всей терминологии (низкая частотность отдельных терминов или разные перевода для одних и тех же терминов, что приводит к нестабильности перевода нейронной моделью).

Преимущества технологии настройки на глоссариях PROMT

- ✓ Добавление глоссария в систему в ручной и автоматическом режимах
- ✓ Возможность редактирования (удаления) отдельных терминов
- ✓ Простой интерфейс, не требуется специальных знаний
- ✓ Учет морфологии в перевода с и на русский с английского и немецкого - не требуется дополнительного постредактирования

PROMT

СПАСИБО ЗА ВНИМАНИЕ!

julia.epiphantseva@promt.ru
Skype: Julia Epiphantseva
Tel., whatsapp: +7 911 2103861