



PROMT ARTIFICIAL INTELLIGENCE TECHNOLOGIES

Настройка NMT на глоссариях

Ожидания и реальность

Юлия Епифанцева

Машинный перевод в бизнесе 2



Способы повышения точности NMT

- **Обучение на параллельных данных**

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

- **Настройка с помощью глоссариев**

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

Актуальность настройки на глоссариях

ПРОБЛЕМА

Обучение нейронной модели на данных, предоставленных заказчиком – самый эффективный способ настройки. Однако для обучения требуется **100+ тыс. сегментов** – такими корпусами **обладает незначительное число компаний**.

РЕШЕНИЕ

Терминологические **глоссарии** или компетенция сотрудников, позволяющая определить правильный перевод того или иного термина, **есть в распоряжении большинства компаний**.

Актуальна задача настройки нейронного перевода через интеграцию глоссария заказчика или прямого редактирования словаря!

PROMT Neural Dictionary – технология нейронного словаря

PROMT Neural Dictionary – специальная технология нейронного словаря, реализованная в нейронных продуктах системы перевода PROMT 21 и доступная для Windows- и Linux-систем.

SmartND

- учитывает контекст
- термины в нужном роде, числе и падеже
- английский ↔ русский, английский ↔ французский, английский < = немецкий

SimpleND

- термины в той же форме, что и в словаре
- 10+ языков, 45+ языковых пар
- успех во многом определяется качеством работы NMT-модели

Требования и возможности

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

SmartND: примеры применения

ИСХОДНЫЙ ТЕКСТ

As a **negative control compound**, the vehicle DMSO was added in the same manner.

Журнал Forbes опубликовал свой ежегодный **рейтинг 100 влиятельных женщин мира** (World's 100 Most Powerful Women), куда в 2011 году было включено немало представительниц ИТ-бизнеса.

БЕЗ СЛОВАРЯ

В качестве **соединения отрицательного контроля** таким же образом добавляли Наполнитель ДМ СО.

Forbes magazine published its annual **ranking** of the **100 most influential women in the world** (World's 100 Most Powerful Women), **where** many IT business members **were included** in 2011.

СО СЛОВАРЕМ

В качестве **негативной регуляции** таким же образом добавляли Наполнитель ДМ СО.

Forbes published its annual **rating** of the **World's 100 Most Powerful Women**, which included many IT-Business members in 2011.

SimpleND: примеры применения

ИСХОДНЫЙ ТЕКСТ

Baugruppenfreigabe für
Verkabelung erteilt

Kunden am Standort FRD
Funkbus/persönlich

БЕЗ СЛОВАРЯ


Cabling assembly release granted

Customers at FRD
Funkbus/personal

СО СЛОВАРЕМ

Assembly release granted for
wiring

Customers at the location FRD
radio bus/personal



**НАСТРОЙКА
NMT ЧЕРЕЗ
ГЛОССАРИЙ НА
ПРИМЕРЕ
КЕЙСОВ**

Глоссарная настройка в разных отраслях

ОТРАСЛЬ	ЯЗЫКОВАЯ ПАРА	ДАННЫЕ
Нефть и газ	ER	Generic NMT + глоссарий
Машиностроение	ER	Tuning NMT + глоссарий
Финансы	ER	Tuning NMT + глоссарий

Результат работы с данными: Нефтегазовая отрасль

Объем глоссария - 3 610 строк

Процент ввода – 66,8%

Материал для тестирования – **реальные документы**

Количество предложений в тестовой части – 8 000+

Количество предложений, перевод которых изменился со словарем по глоссарию – 1 375

Перевод изменился для 17% предложений

Результат работы с данными: Машиностроение

Объем глоссария - 7 591 строка

Процент ввода – 68%

Материал для тестирования – **реальные документы**

Количество предложений в тестовой части – 1 818

Количество предложений, перевод которых изменился со словарем по глоссарию – 355

Перевод изменился для 19,5% предложений

Результат работы с данными: Финансы

Объем глоссария - 1 937 строк

Процент ввода – 94%

Материал для тестирования – строки из ТМ

Количество предложений в тестовой части – 401

Количество предложений, перевод которых изменился со словарем по глоссарию – 31

Перевод изменился для 7,7% предложений

Особенности корпоративных глоссариев

Избыточная или вредная информация для NMT: частотные однословные ключи с общелексическим или узкоспециализированным переводом

English

Russian

Accident

Авария

Vessel

Сосуд

Особенности корпоративных глоссариев

Строки, содержащие скобки, слэши, амперсаны, проценты и т.д.:

English

Air nozzle (spray nozzle)



air nozzle

spray nozzle

Russian

Распылительный штуцер

распылительный штуцер

распылительный штуцер

Особенности корпоративных глоссариев

Строки, содержащие скобки, слэши, амперсаны, проценты и т.д.:

English

Analysis (structural)



structural analysis

Russian

Расчет (при проектирование конструкций)

расчет при проектировании конструкций

Особенности корпоративных глоссариев

Другие случаи:

- Ключ или перевод не являются существительными
- Строки с одинаковым ключом, но отличающимися переводами
- Ключ или перевод оканчивается на предлог (issued for)
- Некорректные символы (амперсанд, другие неалфавитные символы)
- Перевод содержит точку (маркер справочной информации)
- Ключ содержит менее 3 символов
- Кириллические символы в английской части
- Противоречивая информация с т. зрения разделителей перевода (запятая, точка с запятой)
- Строки с прописными буквами

Особенности корпоративных глоссариев

Противоречивая структура исходного глоссария → глоссарий может не коррелировать с переводами из параллельных данных

Оригинал

Модель/ Перевод в ТМ

Перевод с глоссарием

VECM

VECM

Векторная модель коррекции ошибок

Participant

Участник

Субъект

Особенности корпоративного глоссария

Выводы

- Глоссарий всегда содержит избыточную информацию, так как часто его цель - это справочная информация
- Некоторые особенности глоссариев могут быть устранены автоматической прочисткой некорректных статей, другие требуют ручного редактирования
- Объем NMT словаря часто меньше объема исходного глоссария

Где глоссарий помогает

Имена собственные

English

American Industrial Hygiene Association

American Society of Heating, Refrigeration and Air Conditioning Engineers

Russian

Американская Ассоциация
Промышленной Гигиены

Американское общество
инженеров по системам
обогрева, охлаждения и
кондиционирования воздуха

Где глоссарий помогает

Перевод аббревиатур

English

HVAC

AMUR GCC LLC

EMS

DCS

Russian

ОВКВ

ООО «Амурский ГХК»

СНОС

РСУ

Где глоссарий помогает

Уточнение перевода

English

Derrick

Driven end

Russian

Буровая вышка

Приводной конец

Оценка качества перевода

Статистический метод (BLEU)

Generic NMT

Tuning NMT

Tuning NMT + глоссарий

42,4

53,49

51,67

30,41

38,89

38,65

Выводы

- Корпоративный глоссарий может быть эффективным инструментом настройки NMT
- Для качественного результата рекомендуется выделять из корпоративного глоссария потенциально полезную информацию (аббревиатуры, имена собственные, узкоспециализированные частотные словосочетания и т.д.), которые не переводятся корректно с универсальной или тюнингованной NMT моделью
- Для дальнейшего развития словаря NMT рекомендуется использовать механизмы ручного пополнения словаря



**Спасибо
за внимание!**

Julia.Epiphantseva@prompt.ru

+7 (495) 580 48 48

prompt.ru