Нейросетевой МП: ВОПРОСЫ КАСТОМИЗАЦИИ

Women in Localization Russia
Custom MT

Юлия Епифанцева
PROMT

Технологии машинного перевода

НОВЫЕ ТЕХНОЛОГИИ - НОВЫЕ ВОЗМОЖНОСТИ И ТРЕБОВАНИЯ

Rule-Based MT с 1950 гг.

Машинный акцент в переводе, предсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ СЛОВАРИ

Statistical MT с 2000 гг.

Более гладкий перевод чем в RBMT, но непредсказуемый результат

ОБУЧЕНИЕ ЧЕРЕЗ ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ

Neural MT с 2015 гг.

Точный и гладкий перевод, результат на уровне профессионального переводчика

ОБУЧЕНИЕ ЧЕРЕЗ ПАРАЛЛЕЛЬНЫЕ ДАННЫЕ И СЛОВАРИ (ГЛОССАРИИ)

Generic Neural MT outperforms customized RBMT & SMT

PROMT

Целесообразность настройки

- Стартовое качество не удовлетворяет требованиям решаемой задачи. Требуется большой объем постредактирования или постредактирование без настройки неэффективно.

- Есть данные для обучения достаточного объема и качества (параллельные данные и/или глоссарии)

- Есть инфраструктура для обучения, отвечающая всем требованиям, в том числе и требованиям по безопасности (гарантия конфиденциальности данных)

Generic PROMT,Google, Yandex, DeepL

Экспертная оценка качества перевода:

- Нет или мало ошибок в переводе терминологии, названий, аббревиатур?
- Корректная структура предложений? Верная передача синтаксических конструкций?
- Нет ошибок в переводе личных местоимений (мужской/женский род) или соблюдении формального/ неформального стиля?
- Объем постредактирования?

Почему generic NMT ошибается в переводе терминологии?

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

Общая модель всегда будет отдавать предпочтение общим (общелексическим) переводам и не сможет корректно переводить специализированные термины. Аналогично и с синтаксическими конструкциями - NMT модель будет воспроизводить в переводе тот синтаксис, на котором она обучалась.

PROMT

Методы обучения

Дополнительное обучение на обучающей выборке

- Большой объем параллельных данных (для обучения и тестирования)
- Требуются аппаратные ресурсы и ПО (cloud/inhouse)
- Требуются специальные знания
- Высокая точность настройки на определенный стиль и терминологию

Коррекция перевода отдельных терминов

- Любой объем данных (от одного слова до десятков тысяч терминов)
- Не требуются специальные знания (*зависит от поставщика)
- Высокая точность настройки на определенную терминологию

Параллельные данные для обучения
Требования

Translation Memories

- Базы переводов клиентов, создаваемые в CAT инструментах (выровнены по предложению), tmx формат
- Тематическая однородность с переводимыми данными
- (!) Часть данных будет отбракована из-за «технических» ошибок (до 30% данных могут оказаться непригодными)

Тексты и их переводы

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

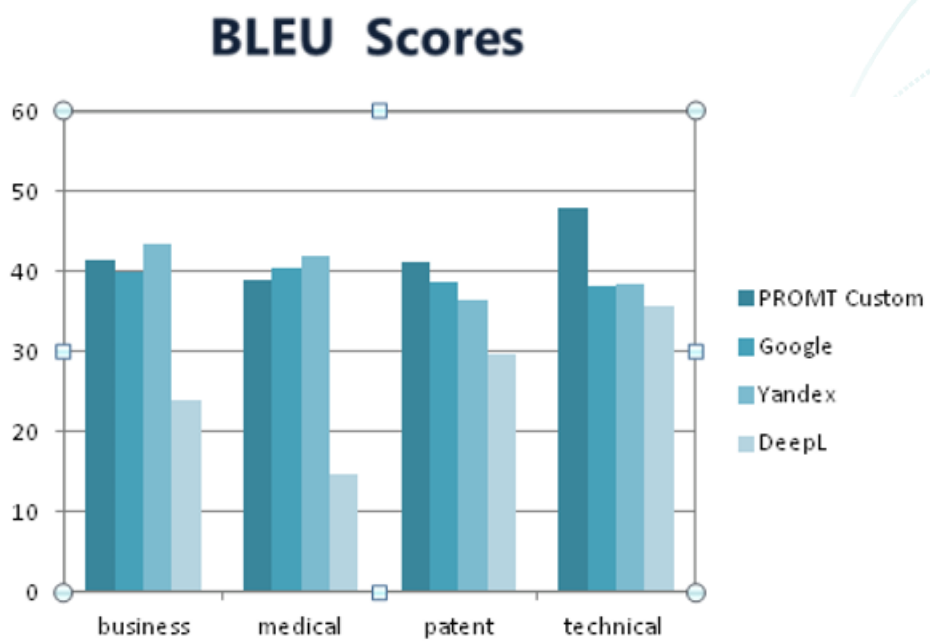
Влияние объема данных на рост качества
Эксперимент PROMT NMT 2020*

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

Customized PROMT & Google, Yandex, DeepL

Кастомизированная модель:

- Более точный перевод терминологии и верная передача синтаксических структур
- Качество перевода зависит от тематики и тренировочных данных
- Уменьшение объема постредактирования



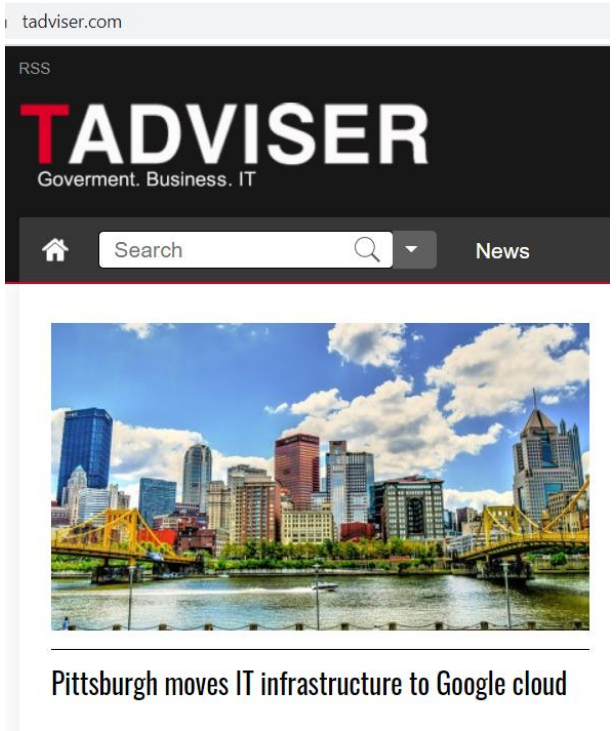
Коррекция переводов терминов

- Любой объем данных - от отдельных слов до десятков тысяч терминов
- Требования к структуре глоссариев, например, 1 ключ - 1 перевод, формат
- Поддержка только существительных или всех основных частей речи
- Используются приложения с пользовательским интерфейсом (PROMT, Systran) или глоссарии подключаются через API (Google Cloud Translation)
- Автоматическое определение морфологии

Локализация сайтов Tadviser.com

- Generic PROMT NMT, русско-английский

- Запланирована настройка через словарь для коррекции перевода имен собственных (названий компаний, подразделений и ведомств, англицизмов в русских текстах).



Специализированная модель Техническая

- Custom PROMT NMT, англо-русский
- Оценка качества перевода (BLEU scores)

  До настройки: 39,3
  После настройки: 47, 93

- Объем данных для настройки: первоначальный объем – 8 млн параллельных сегментов (5,5 млн после фильтрации и валидации)

PROMT

Спасибо за внимание!

julia.epiphantseva@promt.ru

+7 (495) 580 48 48

promt.ru

PROMT