

Трудные вопросы кастомизации машинного перевода

Юлия Епифанцева, PROMT



О компании

ООО «ПРОМТ»

- Отечественный разработчик ПО в области ИИ
- Лучшие эксперты в области МП в России с 1991г.
- Один из ведущих вендоров МП в мире
- Уникальная команда разработчиков – более 40 программистов и лингвистов
- Все продукты компании внесены в Единый реестр отечественного ПО

Клиенты

amadeus



SIEMENS



ЛУК LUKOIL



SpanishD!ct



NORNICKEL
ENSURING RELIABILITY



НИЖФАРМ
ГРУППА КОМПАНИЙ SAGA



3 фактора успеха при внедрении МП

Технология

Передовые достижения современной науки позволяют получить потрясающий результат машинного перевода



Продукт

Гарантия конфиденциальности переводимых данных, гибкая настройка под переводческие бизнес-задачи



Данные

Возможность настройки для перевода терминологии и узкоспециализированных текстов



Международное сотрудничество, службы технической поддержки, локализация документации, интернет-магазинов, сайтов, user-generated content

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.



ОБУЧЕНИЕ НА ПАРАЛЛЕЛЬНЫХ ДАННЫХ

Зачем фильтровать параллельные данные

Объем

Чем больше текстов используется для настройки, тем выше качество (исследования Koehn, Knowles).

Качество

NMT более чувствителен к качеству данных, чем SMT: модель NMT быстро запоминает ошибки и повторяет их в переводе.

Типичные проблемы

Непараллельные строки, не до конца переведенные предложения, сегменты на других языках, слишком длинные строки.

Данные для обучения модели NMT должны быть отфильтрованы и почищены.

Настройка выполняется только на качественных параллельных данных

Оценка BLEU на «чистых»/«грязных» данных для BE>RU модели

КОРПУС	BE > RU на «грязных» данных	BE > RU на «чистых» данных
dev	65,85	69,11
belta	71,59	77,44
sputnik	80,08	86,19

Типы данных

Параллельные данные

- Данные *плохого* качества: сайты, субтитры и т.д.
- Данные *среднего* качества: юридические и правовые данные ООН/Евросоюза
- Данные *хорошего* качества: клиентские данные*, небольшие спец. корпуса, WMT dev/test sets

*Качество клиентских данных

Анализ данных на предмет систематических ошибок, проблемы «человеческого фактора»

Моноязычные данные

- Корпуса новостей
- Wikipedia
- Сайты

Отбор сегментов

- Длина строки более 10 символов
- От 1 до 100 токенов

МП моноязычных корпусов →
синтетические параллельные данные

Схема фильтрации

1. Базовая фильтрация
2. Прочистка по языкам
3. Прочистка с помощью bicleaner
4. Прочистка с помощью nmt-моделей
5. Прочистка с помощью языковых моделей
6. Прочистка URLs и смайликов
7. Валидация

1. Базовая прочистка

ДО ТОКЕНИЗАЦИИ

Инвесторы говорят о «зелёных побегах» выхода из кризиса
(спад уже практически достиг своего предела).

ПОСЛЕ ТОКЕНИЗАЦИИ

Инвесторы говорят о «зелёных побегах»
выхода из кризиса
(спад уже практически достиг своего предела).

- Удаляются дубликаты
- Удаляются тэги

 Текст
 (исходная строка)
 Текст (исправленная строка)
- Удаляется одинаковые source/target строки
- Удаляются слишком длинные и слишком короткие строки
- Удаляются строки, в которых меньше 30% алфавитных символов
- Нормализуются апострофы и кавычки и т.д.

2. Прочистка по языкам

ИСХОДНАЯ СТРОКА	ЯЗЫК, ОПРЕДЕЛЕННЫЙ АВТОМАТИЧЕСКИ
Administration – from 8:00 a.m. to 5:10 p.m.	en
Type of propeller	en
All-Russian Classifier of Territories of Municipal Formations	en
Устанавливая новые цели, мы ориентируемся на лучшие достижения	ru
Обучение персонала	ru
Генеральный директор	ru
About	Unknown (строка слишком короткая)
+7 (888) 201-01-10	Unknown (в строке нет текста)

Родственные языки

- Романские языки: es, fr, ca, it, pt, ro, gl
- Славянский кириллические: ru, uk, mk, bg, sr, be
- Славянские на латинице: cs, sk, sl, hr, sr

3. bicleaner

Удаление неполных, непараллельных и замусоренных строк

Оценка вероятности, что две строки являются переводами друг друга:

- Оценка по шкале от 0 до 1
- 0-0,3 => непараллельные данные
- 0,3-0,5 => оборванные строки, строки с неполным, нетипичным переводом
- 1 => идеальная параллельность

!! Могут отбраковываться строки с обилием терминов (технических и медицинских переводов)

4. nmt-модели

Оценка с помощью nmt-модели, насколько source и target соответствуют друг другу

Source	Target	Оценка
No, I'm gonna mime.	Нет, я сбацаю пантомиму.	-4.05
Does her pipe count?	Даже детская дудочка?	-4.04
Sleeping foxes catch no poultry	Под лежащий камень вода не течет	-3.73
Your adopted son, Brian.	Мы знаем, что он сын Патрисии.	-3.61
For monsters roam in Albion.	Чудовищ полон Альбион.	-2.19
And then this one.	И вот что в результате.	-2.18
Please rephrase your question.	Прошу вас перефразировать вопрос.	-0.48
Eurasian Development Bank.	Евразийский банк развития	-0.0511756

5. Языковые модели

Удаление нечастотных переводов
и очень похожих дубликатов

- Обученная на качественных данных языковая модель
- Проверка корпуса по модели
- Сортировка и удаление сегментов с низкой оценкой
- Удаление похожих дубликатов

Resolution 192/1444

Resolution 192/1445

Resolution 192/1446

Resolution 192/1447

6. URL и смайлики

Удаление смайликов, эмодзи, URL,
#хэштеги, @упоминания и т.д.

СТРОКА ДО ПРОЧИСТКИ

Preprocessor is #awesome

Preprocessor is awesome
<https://github.com/s/preprocessor>

RT @ptwist Preprocessor is awesome :)

СТРОКА ПОСЛЕ ПРОЧИСТКИ

Preprocessor is

Preprocessor is awesome

RT Preprocessor is awesome

7. Валидация

Проверка корпуса на специфические ошибки, которые
повлияют на пунктуацию, регистр и т.д.

Результаты фильтрации

Данные *плохого* качества – удаляется 20-50%

	ИЗВЛЕЧЕНО, СТРОК	ОСТАЛОСЬ ПОСЛЕ БАЗОВОЙ ФИЛЬТРАЦИИ, СТРОК	ОСТАЛОСЬ ПОСЛЕ ОПРЕДЕЛЕНИЯ ЯЗЫКА, СТРОК	ОСТАЛОСЬ ПОСЛЕ VISCLEANER, СТРОК	ОСТАЛОСЬ ПОСЛЕ СКОРИНГА NMT- МОДЕЛЮ, СТРОК
Paracrawl	107 518 323	106 783 061	104 351 522	95 705 207	86 134 507
UN	28 808 505	28 349 382	25 805 088	17 088 080	15 350 000

Результаты фильтрации

Данные *среднего* качества - удаляется 10-15%, данные *хорошего* качества – 5-7%

	ИЗВЛЕЧЕНО, СТРОК	ОСТАЛОСЬ ПОСЛЕ БАЗОВОЙ ФИЛЬТРАЦИИ, СТРОК	ОСТАЛОСЬ ПОСЛЕ ОПРЕДЕЛЕНИЯ ЯЗЫКА, СТРОК	ОСТАЛОСЬ ПОСЛЕ VICLEANER, СТРОК	ОСТАЛОСЬ ПОСЛЕ СКОРИНГА NMT- МОДЕЛЮ, СТРОК
EuroParl	1 948 055	1 932 551	1 915 930	1 787 297	1 608 568
NewsCommentary	375 440	373 409	365 510	343 833	329 450

Результаты фильтрации

Клиентские данные

	ИЗВЛЕЧЕНО ИЗ ТМ, СТРОК	ОСТАЛОСЬ ПОСЛЕ БАЗОВОЙ ФИЛЬТРАЦИИ, СТРОК	ОСТАЛОСЬ ПОСЛЕ VICLEANER, СТРОК	ОСТАЛОСЬ ПОСЛЕ СКОРИНГА NMT- МОДЕЛЮ, СТРОК
ТМ "Медицина"	1 979 588	1 643 452	1 607 620	1 602 676
ТМ «Нефть и газ»	916 286	639 907	-	625 182
ТМ "Техника"	471 991	417 096	403 096	397 663

PROMT Training Add-on

Возможности

- Работает вместе с PROMT Neural Translation Server
- Тренировка на стороне клиента
- Данные для тренировки – файлы в формате tmx, UTF-8
- Рекомендуемый объем- от 50 000 сегментов
(все требования [на сайте](#))

Безопасность&конфиденциальность

- Работает офлайн
- Тренировочные данные не доступны третьим лицам

Результат

- Специализированная модель
- Модель подключается в PROMT Neural Translation Server через профиль перевода
- Профиль перевода доступен для перевода текста, документа, сайта, в CAT-инструменте (Memsource, SDL Trados)



НАСТРОЙКА НА ГЛОССАРИЯХ

PROMT Neural Dictionary

PROMT Neural Dictionary – специальная технология нейронного словаря, реализованная в нейронных продуктах системы перевода PROMT 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

PROMT Neural Dictionary

Возможности

- UI в серверных и десктопных продуктах PROMT (*Мой словарь*)
- Поддержка только существительных
- Поддержка слов и словосочетаний
- 1 ключ - 1 перевод
- Автоматический и ручной ввод

Особенности

- Метрика BLEU покажет ухудшения в качестве перевода
- Скорость перевода со словарем может незначительно снизиться

Не рекомендуется для

- Частотных слов с общелексическими переводами
accident – авария
- Частотных слов с узкоспециальным переводом
vessel – судно

Рекомендуется для

- Имен собственных
- Аббревиатур
- Словосочетаний

SmartND: примеры применения

ИСХОДНЫЙ ТЕКСТ

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

It is theorized that, because the powders have narrow particle size distributions, uniform shrinkage would result in anodes produced from these powders.

БЕЗ СЛОВАРЯ

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

Теоретически, поскольку порошки имеют узкие фракционные составы, равномерная усадка приведет к получению анодов из этих порошков.

СО СЛОВАРЕМ

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

Теоретически, поскольку порошки имеют узкое распределение частиц по размерам, равномерная усадка приведет к получению анодов из этих порошков.



МАШИННЫЙ ПЕРЕВОД СЕГОДНЯ И ЗАВТРА

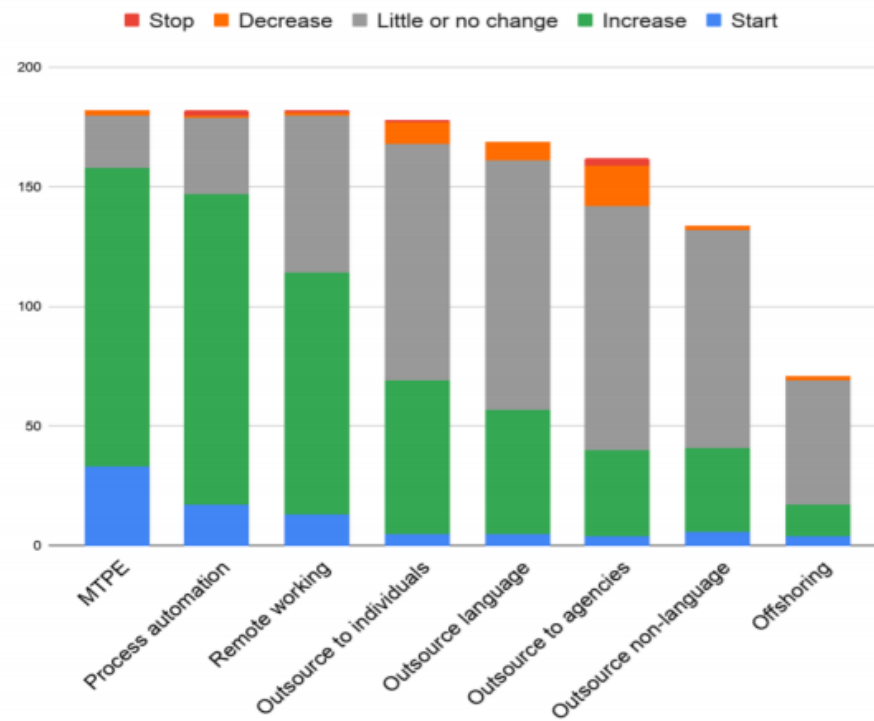
Language Industry Survey 2020

elia. EMT EUATC GALA Lind

Использование МП

- Более востребовано в компаниях, чем у индивидуальных переводчиков
- 158 или 78% опрошенных компаний планируют увеличить или начать использование МП

Operational plans



5 причин задуматься о применении машинного перевода сегодня

- Необходимость перевода с и на иностранные языки никогда не исчезнет
- Технологии будут все больше и больше использоваться для дополнения и расширения работы профессиональных переводчиков
- Нейронный машинный перевод — это инновация и революция в области ИИ
- Актуальный вызов — это не только перевод текста, но и перевод документа
- Вопросы безопасности в ИТ не утратят своей значимости



**Спасибо
за внимание!**

julia.epiphantseva@prompt.ru
+7 (495) 580 48 48
prompt.ru