PROMT

НОВЫЕ ЛИНГВИСТИЧЕСКИЕ
ТЕХНОЛОГИИ

Тюнинг нейронных моделей перевода

Владислав Коваленко
Руководитель отдела машинного обучения

Что такое нейронный машинный перевод?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Это актуальная технология машинного перевода, которая использует нейронные модели, обученные на параллельных корпусах.

В сравнении с rule-based подходом нейронный перевод заметно выигрывает в качестве и гладкости перевода, но также, на первый взгляд, является менее гибким и настраиваемым.

Если нас не устраивает качество перевода лексики, мы не можем просто дать модели словарь и попросить переводить по нему. Но мы можем “дообучить” модель на нужных нам текстах. Это и есть тюнинг.

Что такое тюнинг?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Любую готовую модель можно “дообучить” на новых параллельных данных.

Сам процесс обучения остаётся тем же: алгоритм обновляет веса связей внутри сети, пока модель не научится давать наилучший возможный перевод на предоставленных данных.

Важно: тюнинг – это отдельный процесс, которым управляет пользователь. Модели не умеют обучаться сами по себе.

Нейронные модели в продуктах PROMT

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Можно выделить несколько типов моделей, которые отличаются тем, на каких данных они обучаются:

- Универсальные модели - обучены на десятках миллионов сегментов, покрывающих разные тематические области. Предназначены для перевода любого текста.

- Специализированные - “дообучены” на данных конкретной (относительно широкой) тематики, как то медицина, нефтегаз или ИТ.

- Кастомные - “дообучены” на данных клиента.

Зачем вообще нужен тюнинг?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Универсальные модели – т.н. masters of none. Они дают хороший результат на текстах разных тематик и стилей, но не специализируются на чём-либо конкретном. В частности, часто страдает перевод терминологии:

Вход	Универсальная модель	Специализированная модель	Человеческий перевод
The **field behavior** is very difficult to **match**, especially the late-time water breakthrough.	**Поведение поля** очень трудно **сопоставимо**, особенно поздний прорыв воды.	**Адаптация модели месторождения** представляется затруднительной, особенно **в связи с** поздним прорывом воды.	**Адаптация характеристик месторождения** представляется довольно сложной, особенно **в условиях** прорыва воды на позднем этапе разработки.

Эффект от тюнинга

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Чем больше данных, тем лучше будут результаты тюнинга:

Объём данных	Продолжительность обучения	BLEU на тестовом корпусе
Базовая модель		41.78%
10 тысяч сегментов	6 часов	43.82%
100 тысяч сегментов	5 часов	59.65%
300 тысяч сегментов	17 часов	78.39%

Эффект от тюнинга

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Качество зависит от объёма данных, но:

- прироста в качестве можно добиться и на относительно небольшом объёме
- эффект сильно зависит от однородности текстов клиента

Кейс клиента 1
Прирост в BLEU – 35%, объём данных – ~190 тыс. строк

Кейс клиента 2
Прирост в BLEU – 8%, объём данных – ~3,5 млн. строк

Как правильно выполнить тюнинг?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Если рассматривать тюнинг как задачу машинного обучения, следует рассмотреть четыре важных момента:

- качество обучающих данных;
- использование доменных и внедоменных данных;
- “базовая” модель;
- продолжительность настройки.

Качество обучающих данных

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Во внутренних .tmx часто обнаруживаются не очень адекватные сегменты:

_____/____/____	“____” _____ ____ года
Sitz der Gesellschaft:	Sitz der Gesellschaft:
Employees	A10.17.1
Dear colleagues,	Уважаемые коллеги, добрый день!

Качество обучающих данных

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Эти строки нельзя оставлять в финальном корпусе, на котором будет проводиться тюнинг.

Для их прочистки мы используем алгоритм, который опирается на ряд простых правил. Эти правила учитывают:

- длину сегментов на входе и выходе
- входной и выходной языки
- наличие в тексте букв

и т. д.

Качество обучающих данных

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Объём данных важен для настройки, но их качество ещё важнее. Лучше иметь 50 тысяч хороших сегментов, чем 100 тысяч с мусором.

Примеры статистики по результатам прочистки:

Клиент 1
Исходный корпус – 54 938 сегментов. После прочистки – 45 318.

Клиент 2
Исходный корпус – 309 318 сегментов. После прочистки – 187 405.

Доменные и внедоменные данные

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Простое обучение на доменных текстах приводит к *катастрофическому забыванию*:

Вход	Общая модель	Тюнированная модель
^ Bill Gates **speaks against** DRM..	^ Билл Гейтс **выступает против** DRM..	^ Билл Гейтс **говорит против** DRM..
Since they do not have any medical schools, all medical students have to train outside **Luxembourg** in one of the countries whose medical degrees it recognizes.	Поскольку у них нет медицинских школ, все студенты-медики должны обучаться за пределами **Люксембурга** в одной из стран, чьи медицинские степени она признает.	Поскольку у них нет медицинских школ, все студенты-медики должны тренироваться за пределами **Замбии** в одной из стран, чьи медицинские степени он признает.

Выбор “базовой” модели

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Любой тюнинг является продолжением обучения имеющейся модели на новых данных.

Самым логичным шагом является тюнинг на универсальной модели:
универсальная модель -> доменные данные -> доменная модель

Или же запустить тюнинг уже “дообученной” модели:
универсальная модель -> доменные данные 1 -> доменная модель 1 ->
-> доменные данные 2 -> доменная модель 2

На самом деле “последовательный” тюнинг даст меньший прирост качества, чем простое однократное обучение на всех доступных данных.

Продолжительность тюнинга

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Время обучения модели считается в т.н. “эпохах”.
1 эпоха – время, за которое модель просмотрела весь обучающий корпус 1 раз.

В зависимости от объёма данных 2-5 эпох может быть достаточно для эффективного тюнинга. Обычно это занимает несколько часов.

Продолжительность тюнинга

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

При этом тюнинг может продолжать десятки часов и целые дни. Если не прерывать этот процесс, результат может оказаться ещё лучше.

Объём данных	Продолжительность обучения	BLEU на тестовом корпусе
>300 тысяч сегментов	2 часа	45.98%
>300 тысяч сегментов	17 часов	78.39%

Что делать обычному пользователю?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

Тюнинг – сложный процесс со множеством шагов.

PROMT Neural Training Addon (PNTA) – решение для автоматизации этого процесса, доступное широкому кругу пользователей.

- работа происходит в простом веб-интерфейсе;
- требуется лишь подгрузить .tmx с параллельными данными и выбрать параметры;
- после обучения модель можно сразу подключить в PROMT Neural Translation Server и/или PROMT Translation Factory

Как всё это работает в PNTA?

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

PNTA учитывает все наши наработки по рассмотренным вопросам.

1. Доменные данные проходят автоматическую прочистку внутри PNTA.

2. Перед запуском тюнинга к доменным данным добавляются внедоменные данные в пропорции 1:1.

3. Тюнинг проводится на универсальной модели PROMT, которая поставляется в составе языкового пакета.

4. Минимальная продолжительность тюнинга высчитывается автоматически по формуле. Пользователь также может выбрать опцию более продолжительного тюнинга или задать конкретное кол-во часов.

Требования к PNTA

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

- Процессор класса Intel Core i5 (или выше) или Xeon E3 (или выше) с 16 ядрами минимум
- Графический процессор (GPU) с выделенной видеопамятью 16 Гб и с поддержкой CUDA 12 (требуется установить последнюю версию драйвера GPU для соответствующей ОС)
- Оперативная память: 32 Гб (при объёме тренировочных данных до 1 млн. сегментов)
- Место на диске: 20 Гб

Связь с другими продуктами PROMT

PROMT
НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

PNTA может быть установлен на одном сервере с PROMT Neural Translation Server и PROMT Translation Factory.

- Переводчики работают в PROMT Translation Factory, накапливают память перевода и экспортируют её в .tmx
- Полученный .tmx подгружается в PROMT Neural Training Addon, обучается кастомная модель
- Эта модель подключается к PROMT Neural Translation Server и/или PROMT Translation Factory
- Теперь можно использовать обновлённый машинный перевод

PROMT

НОВЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ

ПО ВОПРОСАМ ТЕСТИРОВАНИЯ И ЗА КОНСУЛЬТАЦИЯМИ ОБРАЩАЙТЕСЬ ПО ЭЛЕКТРОННОЙ ПОЧТЕ

CORPORATE@PROMT.RU