Моделирование неопределенности и устойчивость нейронных сетей в условиях сдвига данных

Выполнил: Кирилл Мажара.

Содержание I

1. Введение

Цель исследования

Цель данного проекта заключается в разработке и исследовании подхода к повышению устойчивости нейронных моделей машинного перевода (NMT) к сдвигу данных за счёт интеграции методов оценки и управления неопределённостью (uncertainty quantification).

Актуальность

1. Сдвиг данных - это норма

В современных системах машинного перевода входной текст может поступать из любых источников и существенно отличаться по стилю, жанру, тематике, лексике или структуре. Это приводит к различным видам сдвига данных (linguistic, domain, covariate, label shift и др.), негативно влияющим на стабильность модели.

2. Высокая цена полного retraining ⇒ нужен селективный fine-tuning

Полное дообучение модели для каждого нового типа данных требует значительных вычислительных ресурсов и времени. Это делает такой подход непрактичным в условиях частого изменения входного распределения.
Кроме того, повторное дообучение может нарушить ранее обученные параметры, что приводит к ухудшению качества на уже освоенных доменах. Это особенно критично при отсутствии механизмов защиты от катастрофического забывания.

Постановка задачи

Рассмотрим задачу машинного перевода в условиях сдвига данных.
Пусть задано множество входных предложений на языке X:

$$x = \{x_1, x_2, \ldots, x_n\}$$

Модель $f(x, \theta)$ с параметрами θ генерирует перевод на язык Y:

$$y = f(x, \theta)$$

Предположим, что данные на входе подчиняются новому распределению:

$$x \sim P_{\text{shifted}}(X) \neq P_{\text{train}}(X)$$

Цель: повысить устойчивость модели f к сдвигу (используя дополнительную модель QE, оценивающую уверенность в переводе).

2. Структура кода

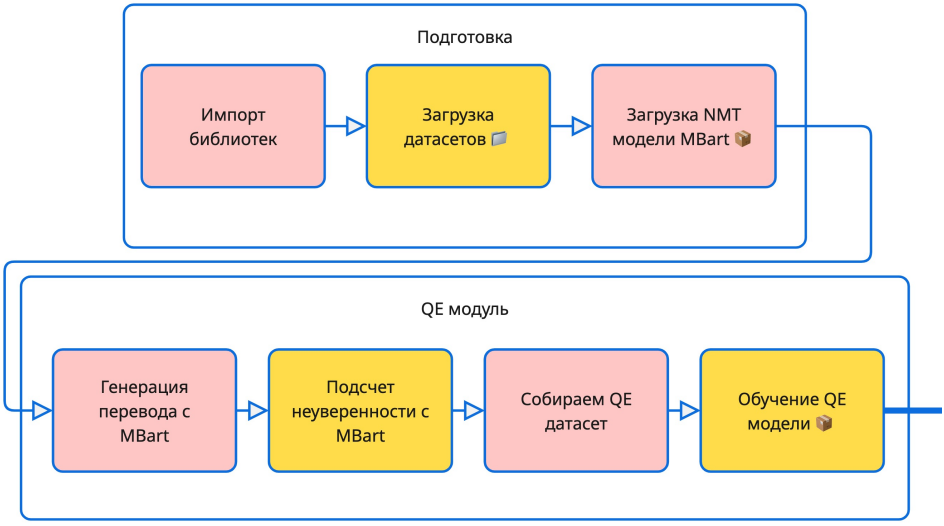
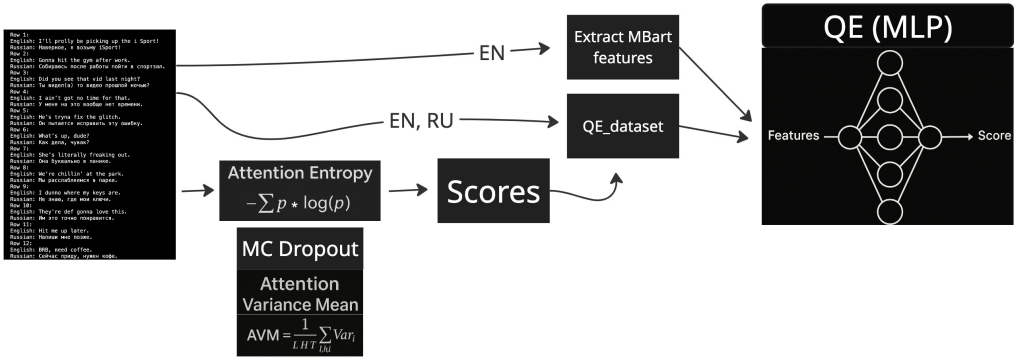


Рис.: Общая логика выполнения кода (часть 1)

Структура кода (часть 1)



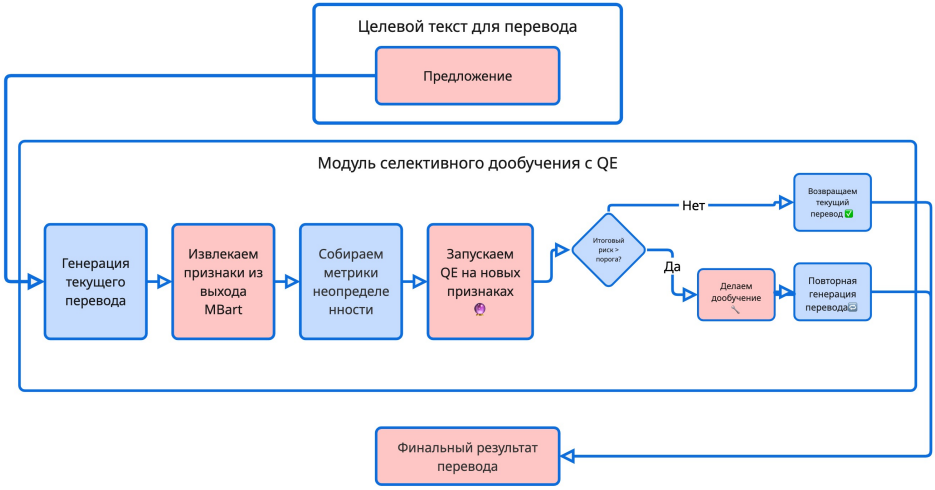


Рис.: Общая логика выполнения кода (часть 2)

Архитектура и используемые модели

Объект исследования

Нейронные модели машинного перевода, функционирующие в условиях сдвига входных данных.

Описание модели MBart

Архитектура — Transformer типа encoder-decoder с механизмом самовнимания. Модель содержит около 610 миллионов параметров;

Задача - генерация перевода с английского на русский

Архитектура QE-модели:

Архитектура - Двухслойный перцептрон регрессор. На вход подаются конкатенация скрытых векторов MBart и трехмерный вектор метрик неопределенности. Выход - риск (скаляр от 0 до 1);

Задача - оценивать неуверенность MBart

Меры качества

BLEU (Bilingual Evaluation Understudy):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log p_i\right)$$

где:

p_i — это точность i-граммы между сгенерированным текстом и исходным текстом,

$n = 4$ — максимальный рассматриваемый порядок n-грамм,

BP — штраф за краткость, наказывающий за слишком короткие переводы.

Формула BP:

$$\text{BP} = \begin{cases} 1, & \text{если } c > r \\ e^{1-\frac{r}{c}}, & \text{если } c \leq r \end{cases}$$

c — длина сгенерированного текста,

r — длина ближайшего по длине эталонного текста.

Меры качества

Precision (точность):

$$\text{Precision} = \frac{\text{число правильных предсказанных токенов}}{\text{общее число предсказанных токенов}}$$

Recall (полнота):

$$\text{Recall} = \frac{\text{число правильных предсказанных токенов}}{\text{общее число токенов в эталоне}}$$

F1-мера — гармоническое среднее:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Предположения

H1. Использование оценки неопределённости на этапе инференса для отбора примеров с высоким риском из датасета с неформальным стилем, и последующего селективного дообучения модели MBart на этих примерах обеспечивает лучшее качество перевода, чем классическиое дообучение на всём целевом корпусе. Улучшение подтверждается ростом мер качества (BLEU, F1, Precision, Recall) на данных с сдвигом.

H2. В тех же условиях предложенный подход позволяет уменьшить объём дообучающих данных и снизить вычислительные затраты, при этом сохраняя или превосходя качество перевода по сравнению с полным дообучением.

Задачи

1. Собрать и подготовить датасеты с контрастными стилями из соревнования Shifts (Reddit <=> UN.en-ru, Shifts MT track);
2. Разработка и реализация подхода на основе модели MBart и вспомогательной модели оценки неопределенности (QE);
3. Проверка эффективности предложенного метода на реальных примерах сдвига данных, используя подготовленные тестовые наборы данны;
4. Анализ и интерпретация полученных результатов, формулирование практических рекомендаций по применению предложенных решений.

3. Детали блока

Reddit-easy

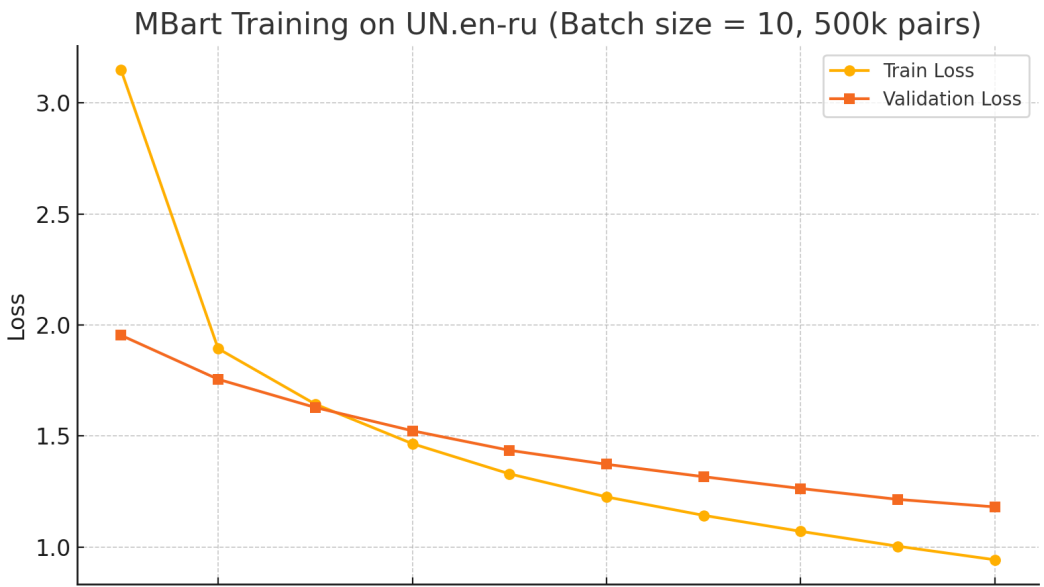
easy = 2347 | hard = 718
Easy samples:

	en	ru	difficulty
107	If you are already doing all of these things, ...	Если ты уже все это делаешь, возможно, твое те...	easy
108	Usually it takes me two weeks to really start ...	Обычно с диетой CICO перед тем, как начать зам...	easy
109	Aubameyang blames 'cursed orange juice' for Ga...	Обамеянг винит «проклятый апельсиновый сок» в ...	easy
110	You probably disagree but idc.	Ты, скорее всего, не согласишься, но мне все р...	easy
111	This sub is toxic so I'm not gonns waste my ti...	Эта токсичный подфорум, и я не буду зря терять...	easy
112	I am already simping for her.	С уже ей сочувствую.	easy
113	Glory to small chested women	Слава женщинам с маленькой грудью.	easy
114	I have thin privilege!	У меня есть привилегия худого человека!	easy
116	Regale RGC cinemark CNK imax and Marcus all th...	Regale, RGC, Cinemark, CNK, IMAx и Marcus: во ...	easy
117	AMC is minimally a 20 dollar share.	Минимальная акция AMC — 20 долларов.	easy
118	Assuming that obt is the last wipe (i.e.	Если допустить, что ОБТ будет последним вайпом...	easy
119	no wipe before release) you could treat it as ...	перед релизом вайпа не будет), можно считать О...	easy
120	Yes there is the possibility of poor f2p pract...	Да, есть вероятность использования плохих прак...	easy
122	Barkhouse taking a breather realizing they los...	Barkhouse берут короткую передышку, поняв, что...	easy
124	I'm not a man, I'm a weapon in human form.	Я не человек, я оружие в форме человека.	easy
125	Just unsheathe me and point me at the enemy.	Просто обнажи меня и направь на врага.	easy
126	But blocking payments to crypto has no effect ...	Но блокировка платежей на криптокошельки не вл...	easy
127	If the money in your Revolut account is illega...	Если деньги на твоем счете Revolut получены не...	easy
128	It's just to force you to buy crypto in Revolu...	Это просто для того, чтобы принудить тебя поку...	easy
129	Lmao was Ainge supposed to move the team to ne...	Ржака, разве Эйндж должен был перевести команд...	easy

Reddit-hard

Hard samples:

	en	ru	difficulty
362	Well then, I feel like if we're going to get i...	Ну тогда, я думаю, если уж мы будем в этом раз...	hard
363	Hell, you probably shed more skin cells in the...	Черт, ты, возможно, сбрасываешь больше частиче...	hard
366	Some missionaries once went to my uncles house...	Однажды к дому моего дяди пришли миссионеры и ...	hard
367	"Bold of you to assume" is a bit of a meme, I ...	«Смело предполагать» это отчасти мем, я написа...	hard
369	There had to be braces....	Нужны были брекеты...	hard
373	All <700 square feet.	Все меньше 700 квадратных футов.	hard
374	Aaah let me take my hands of my trigger while ...	Ааа, позволь мне убрать палец со спускового кр...	hard
376	I also have keqing but Idk what to do with her	У меня также есть Кэ Цин, но я не знаю, что с ...	hard
384	Most of these you can watch on kissasian but a...	Большую часть из них можно посмотреть на KissA...	hard
387	During stims so far (day 8 here), I've had a b...	Во время стимуляции на текущий момент (сейчас ...	hard
392	Ask and you shall receive	Просите, и дано будет вам.	hard
397	When I was making my decision, I was 48 years ...	Когда я принимал решение, мне было 48 лет и я ...	hard
399	I had a lot of lost retirement savings time to...	Я потерял много времени, в течение которого мо...	hard
403	I really thank you for this post because I am ...	Реальное спасибо за этот пост, потому что я то...	hard
406	No science will cure us either	Никакая наука нас тоже не вылечит.	hard
410	You need Conan as iron man	Вам нужен Конан в роли Железного человека.	hard
412	Who's quitting?!?!	Кто уходит?!?!	hard
414	Keep driving the price down hedge fucks, I'll ...	Продолжайте снижать цены, засранцы из хедж-фон...	hard
417	But my gems are free earned from the game, I g...	Но мои гемы бесплатные, заработанные в игре, к...	hard
420	Yes, do it.	Да, сделай это.	hard


MBart Training on UN.en-ru (Batch size = 10, 500k pairs)
Train Loss
Validation Loss
Loss
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10

Примеры уточнения с помощью QE (1/2)

```
================================================================================
SRC       : Barry runs in the same spot until time is reversed back to when he warns Bruce with a
TGT       : Барри бежит в одно и то же место до тех пор, пока время не вернется к тому моменту, ко
BASE_MBART: Барри бегает в том же месте, пока время не возвращается к тому, когда он предупреждает
attention_entropy = 2.8771 | qe_uncert = 0.4124 | combined_risk = 2.1377  -->  ADD_TO_BUFFER
BLEU       | base: 0.198  refined: 0.208
Precision    | base: 0.066  refined: 0.034
Recall    | base: 0.066  refined: 0.034
F1 Score   | base: 0.066  refined: 0.034
REFINED_MBART: Барри бежит в том же месте, пока время не возвращается к тому, когда он предупрежда
```

```
================================================================================
SRC       : Nah the costochrondritis album was superior in all ways 🙄🤚.
TGT       : Нет, альбом Costochondritis был лучше во всех отношениях 🙄🤚.
BASE_MBART: Нет, альбом costochrondritis был в любом случае превосходным.
attention_entropy = 2.0408 | qe_uncert = 0.3292 | combined_risk = 1.5273  -->  ADD_TO_BUFFER
BLEU       | base: 0.172  refined: 0.223
Precision    | base: 0.188  refined: 0.176
Recall    | base: 0.188  refined: 0.176
F1 Score   | base: 0.188  refined: 0.176
REFINED_MBART: Нет, альбом costochrondritis был превосходным во всех отношениях [26].  ⇒ UPDATED
```

Примеры уточнения с помощью QE (2/2)

```
================================================================================
SRC       : Why We Sleep - This book is really eye-opening and really makes you evaluate your rel
TGT       : Почему мы спим: это действительно познавательная книга, которая реально позволяет вам
BASE_MBART: Почему мы спим - Эта книга действительно открывает глаза и действительно делает вас о
attention_entropy = 2.3064 | qe_uncert = 0.3051 | combined_risk = 1.7060  -->  ADD_TO_BUFFER
BLEU       | base: 0.129  refined: 0.150
Precision    | base: 0.200  refined: 0.318
Recall    | base: 0.200  refined: 0.318
F1 Score   | base: 0.200  refined: 0.318
REFINED_MBART: Почему мы спим - Эта книга действительно открывает глаза и действительно делает ва
```

```
================================================================================
SRC       : and players.
TGT       : и у игроков.
BASE_MBART: и игроков.
attention_entropy = 1.1632 | qe_uncert = 0.0729 | combined_risk = 0.8361  -->  OK
BLEU       | base: 0.227  refined: 0.227
Precision    | base: 0.333  refined: 0.333
Recall    | base: 0.333  refined: 0.333
F1 Score   | base: 0.333  refined: 0.333
```

Проверка предположений

H1: улучшение качества перевода (BLEU)

Метод	BLEU (Reddit-easy)	BLEU (Reddit-hard)
Без адаптации	0.238	0.224
Baseline	0.266	0.229
Предлагаемый метод	0.288	0.235

H2: снижение затрат на адаптацию

Метод	GPU-часы	Время эпохи	Память (ГБ)
Без адаптации	-	-	4.2
Полный fine-tuning	0.59	11.7 мин	14.7
Предлагаемый метод	0.31	9.4 мин	10.8

4. Выводы

Выводы

Результаты:

Полученные экспериментальные данные доказывают предположения H1 и H2:

H1: оценка неопределённости + селективный fine-tuning ⇒ повышение качества перевода (BLEU);

H2: уменьшение числа примеров и GPU-затрат при сохранении/превышении качества baseline.

Предложен способ оценки неопределённости перевода с помощью MBart и модели QE: перевод → оценка риска → точечная адаптация.

Сокращены вычислительные затраты на 47% (0.31 GPU-ч vs. 0.59 GPU-ч у полного fine-tuning)

Предлагаемый метод опережает baseline на 7.5% BLEU на Reddit-hard (0.235 против 0.229)

Выводы

Вклад:

Научный: стратегия селективного дообучения только на “рискованных” примерах. Мы научились распознавать полезные примеры.

Практический: масштабируемость, применимость в системах continual learning и онлайн-перевода.

Ограничения и перспективы:

Зависимость от выбора порога риска в QE-модели.

Требуется доступ к референсным переводам (ограничивает автономность).

Одна языковая пара, одна модель $\Rightarrow$ требует расширения.

Контактная информация

Автор: Кирилл Махара

Email: kiraman0403@gmail.com

GitHub: https://github.com/Kiruhins

Telegram: @kiruhins1