ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»
Институт компьютерных наук и технологий

Отчет о прохождении производственной (преддипломной) практики

Фурман Владислав Константинович

(Ф.И.О. обучающегося)

2 курс, 3540203/00101

(номер курса обучения и учебной группы)

02.04.03 Математическое обеспечение и администрирование информационных систем

(направление подготовки (код и наименование))

Место прохождения практики: ФГАОУ ВО «СПбПУ», ИКНиТ, ВШИИ,

(указывается наименование профильной организации или наименование структурного подразделения

г. Санкт-Петербург, ул. Обручевых, д. 1, лит. В

ФГАОУ ВО «СПбПУ», фактический адрес)

Сроки практики: с 20.04.2022 по 20.05.2022

Руководитель практической подготовки от ФГАОУ ВО «СПбПУ»: Белых Игорь Николаевич, к.ф.-м.н., доцент ВШИИ

(*Ф.И.О., уч. степень, должность)*

Консультант практической подготовки от ФГАОУ ВО «СПбПУ»: Пак Вадим Геннадьевич, к.ф.-м.н., доцент ВШИИ

(*Ф.И.О., уч. степень, должность)*

Руководитель практической подготовки от профильной организации: нет

Оценка:

Руководитель практической подготовки
от ФГАОУ ВО «СПбПУ»: Белых И.Н.

Консультант практической подготовки
от ФГАОУ ВО «СПбПУ»: Пак В.Г.

Руководитель практической подготовки
от профильной организации:

Обучающийся: Фурман В.К.

Дата:

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

С помощью естественного языка можно выразить любую мысль, любую идею. Любое изображение или звук можно описать словами. Текст является всеобъемлющим средством передачи информации. Что означает, что обработка текстов на естественных языках (Natural Language Processing, NLP) является крайне важной и актуальной проблемой.

Существует большое множество задач в области обработки текстов, например:

- перевод с одного языка на другой (например, перевод с русского на японский или обратно);
- или же монолингвистический перевод (перевод с языка в него же), как, например, упрощение текстов (понижение сложности слов, выражений, грамматики, сохраняя при этом исходный смысл текста);
- классификация текстов (положительный или отрицательный отзыв, фильтрация спама и т. д.);
- генерация текстов (например, из заданного заголовка сгенерировать статью);
- реферирование текстов (из большого по объёму документа или набора документов выделить ёмкую основную мысль);

Для японского и китайского языков особый интерес представляет задача упрощения текстов, так как эти языки используют иероглифическую письменность, где для чтения текстов нужно знать чтение и значение отдельных иероглифов (в японском языке большинство иероглифов имеют несколько чтений, порой даже больше 10). Это может значительно сузить круг возможных читателей какого-либо текста — дети изучают иероглифы, начиная с первого класса школы и до самого выпуска. То же касается и иностранцев, имеющих довольно ограниченное знание иероглифов. Причём даже взрослые японцы и китайцы могут испытывать трудности с иероглифами, особенно связанные с юридическими документами. Количество иероглифов довольно высоко, в среднем, взрослый японец знает порядка 2 000 иероглифов, взрослый китаец — 8 000 (хотя самих иероглифов значительно больше — не менее 80 000, — но большинство из них используются крайне редко). Поэтому есть высокая потребность в упрощении текстов для увеличения количества их потенциальных читателей.

Более того, упрощение текстов может повысить эффективность других задач NLP, как, например, реферирование, извлечение информации, машинный перевод и т. д.

Предыдущие этапы состояли из:

- обзора предметной области, включающего в себя рассмотрение особенностей обработки японского языка, существующих решений в области упрощения текстов на японском языке, а также примеров их применения;
- теоретического обзора существующих решений, где подробно была рассмотрена модель Transformer и её внутреннее устройство;
- разработки модели с архитектурой Transformer, решающей поставленную задачу упрощения текстов на японском языке, а также сервера и пользовательского веб-приложения.

Целью данной преддипломной практики является завершение работы над системой упрощения текстов на японском языке.

Для достижения данной цели необходимо выполнить следующие задачи:

- проведение заключительных экспериментов с разработанной моделью и её модификациями;
- сравнение полученных метрик BLEU и SARI, а также проверка гипотезы об их улучшении;
- анализ упрощения конкретных предложений на японском языке изначальной и улучшенной моделью.

ГЛАВА 1. МОДИФИКАЦИИ РАЗРАБОТАННОЙ МОДЕЛИ

1.1. Недостатки разработанной модели

Разработанная модель обладает следующими недостатками:

– плохо справляется с большими предложениями[1],
– имеет относительно небольшой «словарный запас».

Обе проблемы вызваны довольно маленьким корпусом (50 000 предложений для такой задачи — крайне малое количество). Самым простым и очевидным решением в такой ситуации было бы взять больший корпус, однако в открытом доступе он попросту отсутствует. Поэтому необходимо искать решение проблем в условиях очень небольшого корпуса.

1.2. Варианты модификации модели

Как мы уже говорили ранее, encoder в Transformer'е можно предобучить, чтобы он лучше кодировал входные последовательности слов предложений. Сделать это можно, например, следующим образом:

– взять корпус из предложений на японском языке (к примеру, предложения с Википедии[2]), сопоставить каждое предложение самому себе (то есть упрощение будет вестись в исходные предложения);
– обучить таким образом модель;
– получить encoder, который имеет какое-то представление о японском языке.

Может возникнуть вопрос: как же мы улучшим модель, если будем обучаться на корпусе, в котором никакого упрощения совсем нет? Секрет кроется в том, encoder не отвечает за само упрощение — он лишь кодирует предложения в матрицы чисел. Поэтому мы можем взять этот encoder и дальше уже обучать изначальную модель с его внедрением. В данной работе мы попробуем использовать 2 стратегии внедрения encoder'а в изначальную модель:

---

[1] На самом деле, эта проблема частично может быть решена разбиением предложения по запятым и отдельному упрощению каждой части, однако лучше, конечно, было бы иметь решение, способное справляться с большими предложениями.

[2] Корпус с предложениями для предобучения модели может быть найден в репозитории модели [1] — файл `modules/Dataset/wikipediaJp/data/wikipediaJp.csv`

1. взять предобученную модель «как есть» и обучить её на корпусе с упрощёнными предложениями;
2. взять изначальную модель и положить в неё предобученный encoder, сгенерировав остальные коэффициенты.

Стоит также отметить, что просто взяв предобученный encoder и положив его в Transformer, мы многого не добьёмся — модель попросту обучится на корпусе с упрощёнными предложениями и пользы от предобученнго encoder'а мы не получим. Чтобы этого не произошло мы уменьшим learning rate для слоёв encoder'а (в данной работе — в 5 раз[3]). В PyTorch это можно сделать следующим образом:

```
encoder = []
rest = []
for name, param in transformer.named_parameters():
  if "encoder" in name:
    encoder.append(param)
  else:
    rest.append(param)

optimizer = torch.optim.Adam(
  [{'params': encoder}, {'params': rest}],
  # ... параметры обучения
)
# здесь мы уменьшаем learning rate у encoder'а в 5 раз
optimizer.param_groups[0]['lr'] = LEARNING_RATE / 5
```

Таким образом мы значительно ограничиваем возможность изменения параметров encoder'а при обучении. Что позволяет нам воспользоваться преимуществом его предобучения.

[3] Число 5 было найдено подбором: сначала была попытка с learning rate в 10, потом в 2, потом в 4 и, наконец, в 5 — что дало лучший результат для метрик.

ГЛАВА 2. АППРОБАЦИЯ РАЗРАБОТАННОЙ МОДЕЛИ И ЕЁ МОДИФИКАЦИЙ

2.1. Метрики для оценки модели упрощения

Для перевода текстов (в том числе и упрощения) довольно часто используют метрку BLEU. В оригинальной статье Папинени и др. [2] показали наличие корреляции данной метрики с сохранением грамматики и смысла переведённых предложений.

Однако есть и более специфичная метрика, разработанная специально для автоматического упрощения текстов — SARI [3]. Она, по сравнению BLEU, лучше коррелирует с упрощением предложений, а также с сохранением лексической и структурной частей предложений.

Вообще говоря, метрика BLEU не очень хорошо подходит для задачи упрощения текстов [4]. Однако во всех источниках, посвящённых упрощению текстов, рассмотренных в данной работе, приводились обе метрики (BLEU и SARI), поэтому и мы поступим аналогично, но будем отдавать метрике SARI больший приоритет.

2.2. Результаты

Результаты метрик BLUE и SARI для изначальной модели и вариантов её улучшения представлены в табл.2.1.

Таблица 2.1

Метрики полученных моделей

Модель	**BLEU**	**SARI**
Transformer	46,98	64,57
Pretrained Transformer	51,12	67,89
Pretrained Encoder	48,22	65,67

В табл.2.1 введены следующие обозначения:

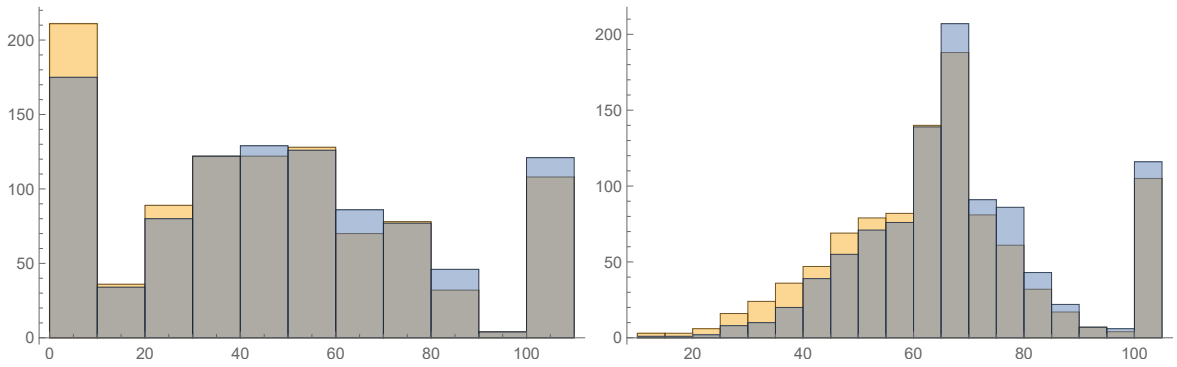
- Transformer — изначальная модель;
- Pretrained Transformer — предобученная модель, которую дообучаем, замедляя обучение encoder'а;
- Pretrained Encoder — изначальная модель, в которой заменяем encoder на предобученный, замедляя его обучение.

По табл.2.1 видно, что наиболее успешной оказалась модель Pretrained Transformer — улучшение по сравнению с изначальной моделью на 4,14% и 3,32% для BLEU и SARI соответственно. Это может говорить о том, что предобучение положительно сказывается и на decoder'е, так как упрощение в основном оставляет предложение в исходном виде, за исключением упрощённых его частей.

2.3. Гистограммы и доверительные интервалы

Рассмотрим гистограммы распределения значений метрик BLUE и SARI для изначальной модели (Transformer) и улучшенной версии (Pretrained Transformer) — см. рис.2.1. Можно увидеть, как в левой части гистограмм (худшие показатели метрик) доминирует первая модель, на правой же части (лучшие показатели метрик) — модифицированная.

Рис.2.1. Сравнение гистограмм метрик BLEU (a) и SARI (b)
жёлтый — Transformer, синий — Pretrained Transformer



a) Метрики BLEU b) Метрики SARI

Это может говорить о том, что улучшение модели действительно положительно влияет на полученные показания метрик, однако для большей уверенности в этой гипотезе вычислим доверительные интервалы для этих значений и посмотрим, не пересекаются ли они. Так как мы имеем большой размер выборки (1 000 элементов), примем гипотезу о нормальном распределении этой выборки и будем считать доверительные интервалы по следующей формуле:

$$CI = \mu \pm p\frac{\sigma}{\sqrt{N}}, \tag{2.1}$$

где

- CI — доверительные интервалы,
- μ — среднее значение выборки,

– p — вероятность попадания истинного значения в доверительные интервалы (в данной работе возьмём значение 0,95),
– σ — среднеквадратическое отклонение выборки,
– N — количество элементов в выборке (в данной работе — 1 000).

Вычислим доверительные интервалы[4]:

$$CI_{\text{Transformer (BLEU)}} = 43{,}85 \pm 0{,}95,\ CI_{\text{Pretrained Transformer (BLEU)}} = 47{,}37 \pm 0{,}94, \quad (2.2)$$

$$CI_{\text{Transformer (SARI)}} = 64{,}57 \pm 0{,}55,\ CI_{\text{Pretrained Transformer (SARI)}} = 67{,}89 \pm 0{,}51. \quad (2.3)$$

В итоге получаем расстояние между нижней границей Pretrained Transformer и верхней границей Transformer в 1,63 и 2,26 для метрик BLEU и SARI соответственно, что может говорить о явном различии истинных значений метрик BLUE и SARI для изначальной и модифицированной моделей. А это, в свою очередь, говорит о наличии значимости внесённых изменений в модель со статистической точки зрения.

2.4. Примеры упрощения предложений изначальной и модифицированной моделями

Рассмотрим примеры предложений из корпуса для тестирования, упрощение которых улучшилось благодаря модификции модели.

В примере на рис.2.3 упрощается слово 内気 на более простое 気が弱い. Однако изначальная модель немного не справилась с упрощением — 弱い (слабая) вместо 気が弱い (скромная)[5].

[4]Заметим, что среднее значение для метрик BLEU отличается от указанных значений в табл.2.1. Связано это с тем, что метрика BLEU вычисляется иным образом для набора предложений, в отличие от единичных предложений, однако разность между значениями при единичных вычислениях ниже, в то время как доверительные интервалы не пересекаются, что говорит о том, что и доверительные интервалы для значений со всеми предложениями не пересекаются. В случае же с SARI разницы никакой нет.

[5]Здесь также происходит упрощение слова ますます (на さらに — более распространённое слово), однако обе модели упростили его одинаково, поэтому не будем заострять на этом внимание.

(1) исходное предложение

kanojo wa uchiki na node masumasu kanojo ga suki da
彼女 は 内気 な ので 、 ますます 彼女 が 好き だ

пер. — она робкая, из-за чего я люблю её ещё больше

(2) изначальная модель

kanojo wa **yowai** node sara ni kanojo ga suki da
彼女 は 弱い ので 、 さら に 彼女 が 好き だ

пер. — она **слабая**, из-за чего я люблю её ещё больше

(3) модифицированная модель (Pretrained Transformer)

kanojo wa **ki ga yowai** node sara ni kanojo ga suki da
彼女 は 気 が 弱い ので 、 さら に 彼女 が 好き だ

пер. — она **скромная**, из-за чего я люблю её ещё больше

Рис.2.3. Пример улучшения упрощения предложения
彼女は内気なので、ますます彼女が好きだ

Рассмотрим ещё один пример — см. рис.2.4. Здесь есть непереводимый на русский язык нюанс упрощения — 我 (литературное «я») → 自分 (обычное «я»)[6]. Заметим также, что в изначальной модели искажается исходный смысл предложения (хотя предложение и упрощается), в модифицированной же модели этот смысл сохраняется (упрощение также производится).

(1) исходное предложение

kare wa okori ni ware wo wasureta
彼 は 怒り に 我 を 忘れた

пер. — он забылся в гневе

(2) изначальная модель

kare wa okotteiru noni **jibun** no **iken wo wasureta**
彼 は 怒っている のに 自分 の 意見 を 忘れた

пер. — он хоть и разозлился, но **забыл своё мнение**

(3) модифицированная модель (Pretrained Transformer)

kare wa okotteiru koto ni **jibun** wo wasureta
彼 は 怒っている こと に 自分 を 忘れた

пер. — он забылся, из-за того что разозлился

Рис.2.4. Пример улучшения упрощения предложения
彼は怒りに我を忘れた

[6]Это именно тот самый случай многообразия японских местоимений (о котором мы говорили в первой главе), каждое из которых несёт свой оттенок и имеет своё место в использовании в японском языке, в русском же языке есть лишь одно слово — «я».

Рассмотрим ещё один пример упрощения — рис.2.5. Здесь можно увидеть упомянутое в первой главе упрощение сложного слова (入場料), состоящего из кандзи, перефразированием более простыми словами (возможно, не самым красивым и лаконичным образом, но значительно более простым). Однако изначальная модель, опять, же исказила смысл предложения, в то время как модель модифицированная передала изначальный посыл.

(1) исходное предложение

nyuu jou ryou wa tada datta
入 場 料 は ただ だった

пер. — вход был бесплатным

(2) изначальная модель

hairu tame no okane wa tada nakatta
入る ため の お金 は ただ なかった

пер. — деньги для входа не были бесплатными

(3) модифицированная модель (Pretrained Transformer)

hairu tame no okane wa zero en datta
入る ため の お金 は 0 円 だった

пер. — денег для входа нужно было 0 йен

Рис.2.5. Пример улучшения упрощения предложения
入場料はただだった

ЗАКЛЮЧЕНИЕ

В данной преддипломной практике была достигнута поставленная цель — завершение работы над системой упрощения текстов на японском языке:

- проведение заключительных экспериментов с разработанной моделью и её модификациями;
- сравнение полученных метрик BLEU и SARI, а также проверка гипотезы об их улучшении;
- анализ упрощения конкретных предложений на японском языке изначальной и улучшенной моделью.

Задачей для следующего этапа является оформление ВКР и её защита.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Репозиторий разработанных модели с сервером. — URL: https://github.com/Ruminat/japanese-simplification (дата обращения: 20.03.2022).

2. Bleu: a Method for Automatic Evaluation of Machine Translation / K. Papineni [и др.] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. — С. 311—318. — DOI 10.3115/1073083.1073135. — URL: https://aclanthology.org/P02-1040.

3. Optimizing Statistical Machine Translation for Text Simplification //. Т. 4. — 2016. — С. 401—415. — URL: https://www.aclweb.org/anthology/Q16-1029.

4. *Sulem E.*, *Abend O.*, *Rappoport A.* BLEU is Not Suitable for the Evaluation of Text Simplification // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium: Association for Computational Linguistics, 2018. — С. 738—744. — DOI 10.18653/v1/D18-1081. — URL: https://aclanthology.org/D18-1081.