Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий

Работа допущена к защите
Зам. директора ВШИСиСТ ИКНТ
_____ А. В. Щукин
«_____» _____ 2022 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ АВТОМАТИЧЕСКАЯ ОБРАБОТКА И ГЕНЕРАЦИЯ ТЕКСТА НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ С ПРИМЕНЕНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

по направлению подготовки 02.04.03.01 Математическое обеспечение и администрирование корпоративных информационных систем
Направленность (профиль) 02.04.03.01_YY Наименование направленности (профиля) образовательной программы

Выполнил
студент гр. 3540203/00101 — В. К. Фурман

Руководитель
доцент каф. ВШИСиСТ,
к. ф.-м. н. — В. Г. Пак

Консультант
по нормоконтролю — В. А. Пархоменко

Санкт-Петербург
2022

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 какой-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.

Целью данной работы является разработка системы автоматического упрощения текстов на японском языке.

Для достижения данной цели необходимо выполнить следующие задачи:

- рассмотреть проблемы обработки текстов на японском языке, а также существующие решения в области упрощения текстов;
- исследовать различные архитектуры нейронных сетей;
- разработать и реализовать описанную систему;
- исследовать качество разработанного решения, а также его эффективность в улучшении других задач NLP.

ГЛАВА 1. ОСОБЕННОСТИ ОБРАБОТКИ ЯПОНСКОГО ЯЗЫКА

1.1. Японская письменность

Японский является очень неординарным языком, сильно отличающимся от европейских, в том числе от русского и английского. На это очень сильно повлиял тот факт, что Япония на протяжении многих веков была закрытой страной для большей части мира. Тем не менее, довольно значимое влияние на японский оказал китайский язык, из которого японцы взяли иероглифы, которые в Японии называют кандзи, но в отличие от китайцев, японцы используют также 2 слоговые азбуки — хирагану и катакану. Примеры японской письменности показаны на рис.1.1.

Катакана: オマエハモウシンデイル
Хирагана: そんなのってないぺこじゃん
Кандзи: 夜露死苦

Рис.1.1. Японская письменность

1.2. Упрощение лексики

Вместе с письменностью в японский язык пришло и немалое количество слов из китайского языка, вообще говоря, практически каждый иероглиф в японском имеет как минимум 2 чтения: онъёми (китайское чтение, хотя часто сильно отличающееся от изначального китайского звучания ввиду особенностей японской фонетики) и кунъёми (японское чтение). Как правило (хоть и не всегда), лексика, пришедшая из китайского языка, значительно труднее исконно японских слов и зачастую упрощение текстов на японском состоит именно в замене таких слов на японские аналоги. На рис.1.2 показан пример упрощения редко встречающегося слова китайского происхождения вполне обычным повседневным лексиконом, состоящим из чисто японских слов и понятному любому школьнику. Как можно заметить, упрощённый результат получился ощутимо длиннее изначального слова.

chishiki houfu
知識豊富 — редкое слово (знаток)

iroiro na koto wo shitteiru
いろいろ な こと を 知っている — простой лексикон (много знающий)

Рис.1.2. Пример упрощения сложного слова

1.3. Морфология в японском

Морфология в японском языке относительно простая — у слов нет ни числа, ни рода, ни падежей, отсутствуют артикли, у глаголов есть всего 2 времени: прошлое и настоящее-будущее. Однако ввиду наличия очень большого количества омонимов (слов, звучащих или пишущихся одинаково, но имеющих разное значение, например, в словаре можно найти более 30 слов с написанием «shi»), могут возникать неоднозначности в письменности. Как правило, в письменности омонимы различаются по иероглифам, используемым в словах, однако в текстах можно встретить эти слова, записанные азбукой, что и создаёт неоднозначности. Такое большое количество омонимов появилось в японском из-за заимствования слов из китайского, где эти омонимы различались по тонам, которые в японском не используются.

1.4. Японская грамматика

В японском, как и в китайском, не используются пробелы, что является существенной проблемой в задаче токенизации (разбиение текста на список токенов — отдельных слов, чисел, дат и т. д.) Существуют готовые решения в области токенизации для обоих языков, однако они сталкиваются с проблемами неоднозначности, которые нельзя решить без глубокого понимания текста и контекста. Пример для японского, где 2 предложения абсолютно идентичны по написанию, но отличаются по смыслу, показан на рис.1.3. Определить, что имеется в виду, можно лишь зная контекст этого предложения.

nande kita no
何で 来た の ？ — **зачем** ты приехал?

nani de kita no
何 で 来た の ？ — **на чём** ты приехал?

Рис.1.3. Пример неоднозначности в японском языке

1.5. Упрощение грамматики

Существуют в японском некоторые грамматические конструкции, сложные для восприятия для не носителей языка, для которых, как правило, существуют более простые аналоги, пример такой конструкции представлен на рис.1.4. Как

правило, такие конструкции встречаются в новостях, различных официальных документах, литературных произведениях и т. п.

wagahai wa neko de aru
吾輩 は 猫 で ある (пер.: «я — кот»)
de aru da
формальная конструкция で ある (можно проще — だ)

Рис.1.4. Пример сложной грамматической конструкции

Вообще говоря, в большинстве случаев они не представляют большой сложности, однако порой могут затруднить понимание текста для людей, нечасто сталкивающимися с подобными текстами, или же теми, кто плохо знает японскую грамматику.

ГЛАВА 2. СУЩЕСТВУЮЩИЕ РЕШЕНИЯ И ДАТАСЕТЫ

2.1. Модель Transformer

Маруяма Т. и Ямамото К. использовали в своём исследовании [9] относительно новую модель Transformer [5]. Они предобучили свою модель на статьях с японской википедии, после чего дообучили (fine-tune) её на небольшом параллельном корпусе, состоящим из 1 100 документов, составленных 40 учителями японского языка [10]. Авторы показали, что в условиях малого количества ресурсов (отсутствия объёмных корпусов для упрощения японских текстов), модель Transformer показывает довольно хорошие результаты — она существенно обходит существующие на сегодняшний день решения в обоих метриках BLEU и SARI, которые обычно используют в задаче упрощения текстов.

2.2. Улучшение упрощения увеличением корпуса с обучением без учителя

Кацута А. и Ямамото К. попробовали создать модель, не требующую параллельного корпуса, то есть их модель может обучаться без учителя [8]. Их подход заключается в создании псевдокорпуса из неразмеченного веб-корпуса, они показали, что расширение такого корпуса ведёт к улучшению результатов упрощения.

2.3. Упрощение новостей

Гото И., Танака Х. и Кумано Т. в своей работе [6] провели исследование в области упрощения новостных текстов. Они сконструировали корпус из новостей с News Web Easy, где новости проходят следующую обработку: сначала они оставляют лишь основную информацию, тем самым укорачивая статью (особенно когда исходные статьи слишком длинные); после чего упрощаются отдельные выражения в предложениях. После чего они вручную разметили этот корпус, обучили модель статистического монолингвистического (из языка в него же) машинного перевода. В итоге они получили вполне неплохие результаты. Однако сравнивать их с результатами других работ затруднительно из-за специфики области их исследования — упрощения новостных текстов.

2.4. JSSS корпус

Такамичи С., Комачи М. и др. составили корпус для упрощения и реферирования японской речи [7]. Корпус содержит проговорённые дикторами тексты, для каждого текста есть таймкоды для синхронизации текста и речи, для упрощения есть параллельные упрощённые тексты, для реферирования, соответственно, приведены рефераты текстов. Тем не менее размер данного корпуса довольно мал — он содержит лишь несколько сотен предложений, — поэтому за основу его брать нельзя, однако его можно попробовать использовать для объединения с другими корпусами.

ГЛАВА 3. ГДЕ ИСПОЛЬЗУЮТ УПРОЩЕНИЕ

3.1. Simple English Wikipedia

Для английского языка существует упрощённая википедия, где статьи вручную переведены в упрощённый вариант английского (Simple English), использующий приблизительно 1 500 одних из наиболее употребляемых английских слов. Simple English основан на Basic English, использующий 850 слов, созданный Чарльзом К. Огденом. На ноябрь 2020 упрощённая википедия содержит более 177 000 статей [3].

3.2. NHK News Web Easy

Чего-то столь же масштабного для японского языка не существует. Есть новостной сайт News Web Easy, на котором выкладывают упрощённые версии новостей NHK (одна из крупнейших японских СМИ) для учеников младшей и средней школ (дети до 15 лет) и иностранцев, проживающих в Японии. Главная страница сайта показана на рис.3.1. Как и на Simple Wikipedia, упрощение новостей на News Web Easy происходит вручную [2].

Рис.3.1. Главная страница News Web Easy

3.3. Текущее положение дел в системах упрощения текстов

Вообще говоря, на сегодняшний пока ещё день не существует достаточно качественной системы упрощения текстов, способной заменить ручной перевод — дела здесь обстоят немногим лучше машинного перевода (из одного языка в другой), что можно объяснить отсутствием качественного и масштабного корпуса для обучения модели упрощения текстов (не только для японского языка, но даже для английского) и довольно высокой сложностью самой задачи, связанной с необходимостью «понимать» текст, что, к сожалению, современный искусственный интеллект сделать пока ещё не в состоянии.

Тем не менее, подобно тому, как сегодня используются системы машинного перевода (например, перевод отдельных слов или перевод текстов с последующими ручными корректировками), могут использоваться и системы упрощения текстов.

ГЛАВА 4. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ РАБОТЫ

4.1. Этапы обработки естественного языка

Как правило, в обработке текстов обычно выделяют следующие этапы [11, с. 9]:

1. Графематический анализ. Здесь осуществляется анализ на уровне символов, в том числе и токенизация, то есть разбиение набора символов (текста) на последовательность отдельных структурированных частей (слово, знак препинания, число, гиперссылка, адрес электронной почты и т. д.).
2. Морфологический анализ. Здесь происходит анализ на уровне слов (не токенов). Стоит выделить следующие процессы, проходящие на этом этапе:
   - Лемматизация — это нахождение нормальной (начальной) формы слова (леммы), к примеру, лемма у слова «сбегать» — «бегать». В японском это относится в основном к глаголам, так как у существительных, как правило, всего одна словоформа.
   - Приписывание граммем. Граммемы — грамматическая характеристика, например, род, падеж, число. Граммемы могут помочь в разрешении неоднозначностей, которые возникают в морфологии.
3. Фрагментационный анализ. Осуществляется на уровне фраз, частей предложений. Этот этап неразрывно связан с синтаксическим анализом, а иногда и вовсе говорят о них, как об одном целом. Сюда, например, может входить обработка причастных или деепричастных оборотов.
4. Синтаксический анализ. Осуществляется на уровне предложений. Здесь, как правило, строится дерево зависимостей одних слов от других, а также исключается морфологическая неоднозначность.
5. Семантический анализ. Осуществляется на уровне всего текста. Самый сложный и неоднозначный из этапов, здесь появляется формальное представление смысла текста, как правило, в виде сементического графа. На сегодняшний день задачи семантического анализа чаще всего решаются нейронными сетями, о чём мы и поговорим далее.

4.2. Нейронные сети

Вкратце, нейронная сеть представляет собой систему соединённых и взаимодействующих между собой простых нейронов. Каждый нейрон получает на вход несколько чисел (входные данные или же выходы других нейронов), суммирует эти числа с определёнными коэффициентами (нахождение оптимальных коэффициентов — обучение нейронной сети), после чего применяет к сумме функцию активации (любую нелинейную функцию) и передаёт результат на вход другому нейрону (или же на выход нейронной сети).

Существует большое множество различных архитектур нейронных сетей (свёрточные, рекуррентные, рекурсивные, графовые и т. д.), однако в данной работе будет сконцентрировано внимание на так называемых трансформерах (Transformer), используемых, как правило, для языковой обработки, в частности для задач перевода и упрощения текстов.

4.3. О модели Transformer

Transformer — относительно новая (2017 г.) архитектура глубоких нейронных сетей, разработанная в Google Brain. Так же, как и рекуррентные нейронные сети (РНС), трансформеры предназначены для обработки последовательностей (к примеру, текста), то есть трансформеры относятся к sequence-to-sequence (seq2seq) моделям. В отличие от РНС, трансформеры не требуют обработки последовательностей по порядку, благодаря чему они распараллеливаются легче, чем РНС, и могут быстрее обучаться.

Используются трансформеры, например, в Яндексе (там его используют для лучшего ранжирования запросов, то есть поиск идёт не только по тексту, как обычной строке, но и по смыслу этого текста), во многих переводчиках (Яндекс, Google, DeepL и т. д.), а также в GPT-3 — самой большой на сегодняшний день модели генерации текстов на английском языке.

Трансформер состоит из кодировщика и декодировщика (encoder & decoder). Кодировщик получает на вход последовательность слов в виде векторов (word2vec). Декодировщик получает на вход часть этой последовательности и выход кодировщика. Кодировщик и декодировщик состоят из слоев. Слои кодировщика последовательно передают результат следующему слою в качестве его входа. Слои

декодировщика последовательно передают результат следующему слою вместе с результатом кодировщика в качестве его входа.

Каждый кодировщик состоит из механизма внимания (attention) (вход из предыдущего слоя) и нейронной сети с прямой связью (вход из механизма внимания). Каждый декодировщик состоит из механизма внимания (вход из предыдущего слоя), механизма внимания к результатам кодирования (вход из механизма внимания и кодировщика) и нейронной сети с прямой связью (вход из механизма внимания).

Каждый механизм внимания параметризован матрицами весов запросов W_Q, весов ключей W_K, весов значений W_V. Для вычисления внимания входного вектора X к вектору Y, вычисляются вектора $Q = W_Q X$, $K = W_K X$, $V = W_V Y$. Эти вектора используются для вычисления результата внимания по формуле (4.1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4.1}$$

4.4. Выбор инструментов

Для обучения модели будет использоваться Python в связке с фреймворком для машинного обучения TensorFlow [4], предоставляющий широкие возможности для реализации нейронных сетей, в том числе, там присутствует поддержка ранее упомянутых трансформеров. Для токенизации будет использоваться библиотека MeCab [1].

Система будет доступна в браузере в виде простого приложения, то есть модель будет обучена на Python, а пользоваться обученной моделью можно будет в любом современном браузере (пользователю ничего не нужно будет устанавливать). Для реализации веб-приложения будет использоваться JavaScript с фреймворком Svelte, позволяющим создавать современные и быстрые SPA-приложения.

ЗАКЛЮЧЕНИЕ

В данной научно-исследовательской работе были поставлены цель и задачи предстоящей магистерской дипломной работы, были рассмотрены следующие её аспекты:

- собенности обработки японского языка — в частности, японская письменность, морфология, грамматика, упрощение лексики с грамматикой с примерами;
- существующие решения и датасеты — были рассмотрены актуальные исследования в данной теме, а также датасеты, которые в них использовались, было решено разрабатывать систему, основанную на модели Transformer, так как она кажется наиболее перспективной;
- «где используется упрощение» — были рассмотрены 2 сервиса, в которых применяют упрощение, о котором говорилось в НИР: Simple English Wikipedia и News Web Easy, была также затронута тема текущего положения систем упрощения в целом.
- теоретическая часть — были рассмотрены основные этапы обработки текстов на естественном языке, было описано, что такое нейронные сети и модель Transformer, был также рассмотрен инструментарий для практической реализации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. MeCab. — URL: https://github.com/taku910/mecab (дата обращения: 21.05.2021).

2. News Web Easy. — URL: https://www3.nhk.or.jp/news/easy/ (дата обращения: 24.11.2020).

3. Simple English Wikipedia. — URL: https://www.wikiwand.com/en/Simple_English_Wikipedia (дата обращения: 24.11.2020).

4. Tensorflow. — URL: https://www.tensorflow.org (дата обращения: 21.05.2021).

5. Attention Is All You Need / A. Vaswani [и др.]. — 2017. — arXiv: 1706.03762 [cs.CL].

6. *Goto I.*, *Tanaka H.*, *Machine Translation Summit XV T. K. of.* Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text // Machine Translation Summit XV. — 2015. — Т. 1. — С. 17—31.

7. JSSS: free Japanese speech corpus for summarization and simplification / S. Takamichi [и др.]. — 2020. — arXiv: 2010.01793 [eess.AS].

8. *Katsuta A.*, *Yamamoto K.* Improving text simplification by corpus expansion with unsupervised learning. — 2019. — DOI 10.1109/IALP48816.2019.9037567.

9. *Maruyama T.*, *Yamamoto K.* Extremely Low Resource Text simplification with Pre-trained Transformer Language Model. — 2019. — DOI 10.1109/IALP48816.2019.9037650.

10. *Moku M.*, *Yamamoto H.* Automatic Easy Japanese Translation for information accessibility of foreigners. — 2012. — URL: https://www.aclweb.org/anthology/W12-5811.

11. *Батура Т. В.* Математическая лингвистика и автоматическая обработка текстов на естественном языке. — Новосибирск: Новосиб. гос. ун-т., 2016. — 166 с.