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: おまえはもうしんでいる
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. 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

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.

ГЛАВА 2. СУЩЕСТВУЮЩИЕ РЕШЕНИЯ И ДАТАСЕТЫ

2.0.1. Модель Transformer

Маруяма Т. и Ямамото К. использовали в своём исследовании [6] относительно новую модель Transformer [3]. 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 [7]. 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.

2.0.2. 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 [5]. Их подход заключается в создании псевдокорпуса из неразмеченного веб-корпуса, они показали, что расширение такого корпуса ведёт улучшению результатов упрощения.

2.0.3. JSSS корпус

Такамичи С., Комачи М. и др. составили корпус для упрощения и реферирования японской речи [4]. 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

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 [2].

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 [1].

Рис.3.1. Главная страница News Web Easy

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.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

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

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

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

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

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

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

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