Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа искусственного интеллекта

Выпускная квалификационная работа магистра

РАЗРАБОТКА И ИССЛЕДОВАНИЕ СИСТЕМЫ АВТОМАТИЧЕСКОГО УПРОЩЕНИЯ ТЕКСТОВ НА ЯПОНСКОМ ЯЗЫКЕ

Направление:
02.04.03 «Математическое обеспечение и администрирование информационных систем»

Выполнил:
студент гр. 3540203/00101
Фурман Владислав Константинович

Научный руководитель:
к. ф.–м. н., доцент ВШИИ
Пак Вадим Геннадьевич

Санкт-Петербург
2022 г.

Цель

Разработка и исследование системы автоматического упрощения текстов на японском языке.

Задачи

- Исследование предметной области.
- Исследование существующих решений и технологий (ИНС).
- Разработка системы упрощения, а также её улучшение.
- Сбор и анализ метрик, обзор упрощения предложений разработанной системой.

Мотивация

В качестве мотивации выступают следующие доводы:

- японский язык очень непростой (порой и для самих японцев), в основном из-за иероглифов (но не только);
- упрощение текстов — это расширение их потенциальных читателей, упрощение понимания смысла текстов;
- с системами упрощения на сегодняшний день всё непросто — их мало и они закрыты (не выходят за рамки статей).

Как упрощать тексты на японском

В японском есть 3 вида письменности:

- две азбуки — хирагана (ひらがな) и катакана (カタカナ);
- иероглифы (кандзи) (漢字).

Как можно упрощать тексты на японском:

- заменять сложные слова (обычно из кандзи) на простые;
- менять формальные грамматические конструкции на разговорные;
- заменять использование некоторых иероглифов на азбуку.

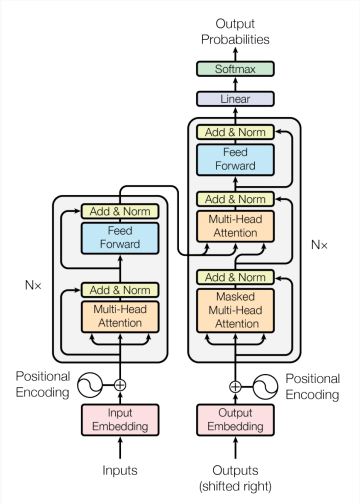
Модели для решения задачи упрощения

- **Рекуррентные нейронные сети (RNN)** (1982 г.)
  - Обработка последовательностей (например, текста).
  - На каждый слой передаётся текущий элемент (слово) + результат предыдущего слоя.
  - Причём есть обратные связи — поэтому рекуррентные.
  - Очень медленные — из-за последовательной природы **нельзя распараллелить**.
- **Долгая краткосрочная память (LTSM)** (1997 г.)
  - Разновиднсть RNN с элементом «забывания».
  - Ещё медленнее.
- **Transformer** (2017 г.) — значительно быстрее за счёт распараллеливания + выше качество.

Архитектура Transformer’a

Состоит из:

- encoder’a — переводит предложение в понятный модели контекст;
- decoder’a — превращает контекст во что-то полезное;
- генератора — генерирует следующее слово в предложении;
- механизма внимания и positional encoding — о них поговорим далее.



Механизм внимания

Механизм внимания может быть представлен формулой:

$$\mathrm{Attention}(Q, K, V) = \underbrace{\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)}_{\text{scores}} V, \tag{1}$$

где

- scores «оценивают» важность элементов (там лежат значения от 0 до 1);
- Q (Query), K (Key), V (Value) — матрицы входных элементов;
- d_k — нижняя размерность одной из этих матриц (длина части embedding'а);
- $\mathrm{softmax}$ — функция нормировки (зажимает значения вектора в отрезок $[0; 1]$, сумма координат становится равной единице).

Positional encoding

Так как в Transformer'е нет ни рекурренции (recurrence), ни свёртки, нам нужно что-то, что будет использовать порядок элементов в последовательности (positional encoding):

$$PE(p, 2i) = \sin\left(\frac{p}{10\,000^{2i/d_{\text{model}}}}\right), \tag{2}$$

$$PE(p, 2i+1) = \cos\left(\frac{p}{10\,000^{(2i+1)/d_{\text{model}}}}\right), \tag{3}$$

где

- p (position) — позиция,
- i (dimension) — размер предложения.

Архитектура системы

Система состоит из:

1. Клиентского приложения.
   Написано на TypeScript + Lit, минималистичный дизайн с формой для ввода предложения на японском, ниже — вывод упрощённого варианта.
2. Сервера.
   Написан на Python + Falcon + MeCab, получает запрос с японским предложением, возвращает токены + упрощённый вариант.
3. Модели ИНС.
   Написана на Python + PyTorch + Spacy + HuggingFace, архитектура Transformer.

Корпус SNOW

Маруяма Т. и Ямамото К. вручную составили корпус из 85 000 предложений с их упрощёнными вариантами.

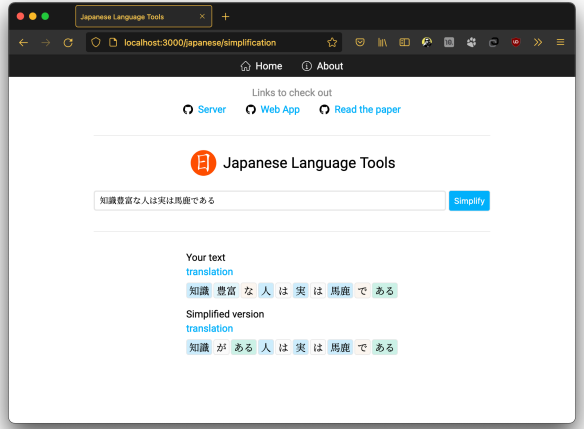
Словарь упрощённых предложений в корпусе составляет лишь 2 000 слов.

Корпус состоит из 2-х частей:

1. SNOW 15: 50 000 предложений (только обучение),
2. SNOW 23: 35 000 предложений (33 000/1 000/1 000 — train/valid/test).

Пользовательское приложение

Минималистичный дизайн: пользователь вводит предложение, нажимает Enter, снизу выводится его упрощённая версия.



Проблемы реализованной системы

Обученная модель обладает следующими недостатками:

- плохо справляется с большими предложениями,
- имеет относительно небольшой «словарный запас».

Решение — предобучить (pretrain) модель на неразмеченном корпусе и дообучить (fine-tune) на корпусе SNOW:

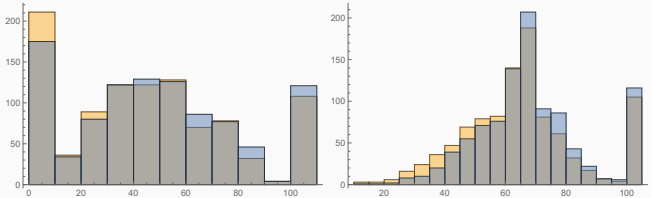
- Pretrained Transformer — предобучение всей модели,
- Pretrained Encoder — предобучение лишь encoder'а.

Метрики

Метрики BLEU и SARI обученных моделей:

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

Сравнение гистограмм метрик BLEU и SARI (Transformer / Pretrained Transformer):



Пример упрощения №1

(1) Исходное предложение:
彼は怒りに我を忘れた
— он забылся в гневе.

(2) Изначальная модель (Transformer):
彼は怒っているのに自分の意見を忘れた
— он хоть и разозлился, но забыл своё мнение.

(3) Модифицированная модель (Pretrained Transformer):
彼は怒っていることに自分を忘れた
— он забылся, из-за того что разозлился.

Интересный момент: « 我 » → « 自分 », — и то, и другое на русском просто «я», но на японском разные оттенки.

Пример упрощения №2

(1) Исходное предложение:
入場料はただだった
— вход был бесплатным.

(2) Изначальная модель (Transformer):
入るためのお金はただなかった
— деньги для входа не были бесплатными.

(3) Модифицированная модель (Pretrained Transformer):
入るためのお金は0円だった
— денег для входа нужно было 0 йен.

Здесь происходит замена сложного слова из кандзи (入場料) на простую фразу (入るためのお金).

Пример упрощения №3

(1) Исходное предложение:
そのスキャンダルはやがてみんなに知れ渡るだろう
— об этом скандале, вероятно, скоро узнают все.

(2) Изначальная модель (Transformer):
その事件を守る事件はやがてみんなに知られるだろう
— скоро об этом событии, защищающем событие, вероятно, узнают все.

(3) Модифицированная модель (Pretrained Transformer):
その悪い話はやがてみんなに知られるだろう
— об этой нехорошей истории скоро, вероятно, все узнают.

Интересный момент: в корпусе представлен менее удачный вариант упрощения данного предложения.
その悪い、知られたくないことは、やがてみんなに報告されるだろう
— Об этом нехорошем деле, о котором никто не хочет знать, скоро всем доложат.

Заключение

Поставленная цель была достигнута — была разработана и исследована система для упрощения текстов на японском языке.

- было проведено исследование предметной области;
- были исследованы существующие решения и технологии (ИНС), в частности — Transformer;
- была разработана система упрощения текстов на японском языке;
- на основе обнаруженных недостатков были внесены улучшения в модель упрощения;
- были собраны и проанализированы метрики BLEU и SARI, а также был проведён обзор упрощения предложений разработанной системой.