Aspect-Based Sentiment Analysis of Japanese Hotel Reviews

Tan Wei Chiong

February 28, 2024

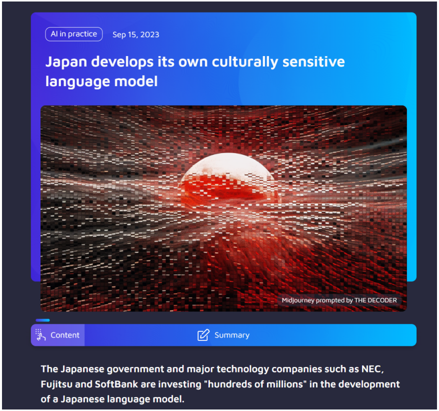
1. Introduction

2. Data Processing

3. Modelling

4. Limitations and Future Work

Motivation



Sep 15, 2023

Japan develops its own culturally sensitive language model

The Japanese government and major technology companies such as NEC, Fujitsu and SoftBank are investing "hundreds of millions" in the development of a Japanese language model.

Figure: (Source: The Decoder. https://the-decoder.com/japan-develops-its-own-culturally-sensitive-language-model/)

Problem Statement

- A hotel chain owner who mainly caters to Japanese tourists wants to find out how good or bad of an experience tourists had during their stay in the hotel.

Problem Statement

- A hotel chain owner who mainly caters to Japanese tourists wants to find out how good or bad of an experience tourists had during their stay in the hotel.
- However, many Japanese do not know enough English to leave a review without using an automatic translator application, and are much more comfortable leaving reviews in their native Japanese.

Problem Statement

- A hotel chain owner who mainly caters to Japanese tourists wants to find out how good or bad of an experience tourists had during their stay in the hotel.
- However, many Japanese do not know enough English to leave a review without using an automatic translator application, and are much more comfortable leaving reviews in their native Japanese.
- As a non-Japanese small business owner, it would be a great help to have a data-driven tool that can help to interpret their true sentiments about your service.

Project Scope

- Disclaimer: This is just a proof of concept.

Project Scope

- Disclaimer: This is just a proof of concept.
- Provide a brief introduction to Japanese NLP methods and tools.

Project Scope

- Disclaimer: This is just a proof of concept.
- Provide a brief introduction to Japanese NLP methods and tools.
- Briefly compare two Japanese sentiment analyzers – oseti and asari.

Data Source

Why The Japanese Website Is Better For Scraping

- Non-dynamic webpage – easier access to relevant information;

Why The Japanese Website Is Better For Scraping

- Non-dynamic webpage – easier access to relevant information;
- Tends to be cluttered, but does not matter much to a scraper.

Why The Japanese Website Is Better For Scraping

- Non-dynamic webpage – easier access to relevant information;
- Tends to be cluttered, but does not matter much to a scraper.
- All reviews need to have the actual review text; the English site does not require that.

Final Dataset

- 7340 reviews total, from 182 hotels across the 6 prefectures of the Tohoku region:
  (Akita, Aomori, Fukushima, Iwate, Miyagi, Yamagata);

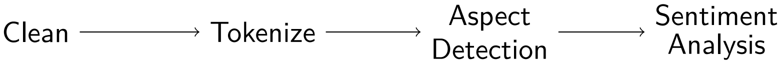
Final Dataset

- 7340 reviews total, from 182 hotels across the 6 prefectures of the Tohoku region:
  (Akita, Aomori, Fukushima, Iwate, Miyagi, Yamagata);
- Reviews without ratings are ignored by the scraper;

Final Dataset

- 7340 reviews total, from 182 hotels across the 6 prefectures of the Tohoku region:
  (Akita, Aomori, Fukushima, Iwate, Miyagi, Yamagata);
- Reviews without ratings are ignored by the scraper;
- 5568 reviews have all scores populated; this is used mainly for performance metrics.

Data Workflow



Tokenization

Let’s try to tokenize a sentence in English!

Tokenization

Let’s try to tokenize a sentence in English!

The reception was very polite.

Tokenization

Let’s try to tokenize a sentence in English!

The reception was very polite.

Easy? What about a Japanese sentence?

Tokenization

Let's try to tokenize a sentence in English!

The reception was very polite.

Easy? What about a Japanese sentence?

フロントはとても丁寧でした。

Japanese Language 101 – Hiragana

フロント は とても 丁寧 でした。

Japanese Language 101 – Hiragana

フロント は とても 丁寧 でした。

- Basic alphabet of Japanese;

Japanese Language 101 – Hiragana

フロント は とても 丁寧 でした。

- Basic alphabet of Japanese;
- Used for particles;

Japanese Language 101 – Hiragana

フロント は とても 丁寧 でした。

- Basic alphabet of Japanese;
- Used for particles;
- Used for words without kanji representation.

Japanese Language 101 – Katakana

フロント は とても 丁寧 でした。

Japanese Language 101 – Katakana

フロント は とても 丁寧 でした。

- Same pronunciations as hiragana, different script;

Japanese Language 101 – Katakana

フロント は とても 丁寧 でした。

- Same pronunciations as hiragana, different script;
- Used for loan words.

Japanese Language 101 – Kanji

フロント は とても 丁寧 でした。

Japanese Language 101 – Kanji

フロント は とても 丁寧 でした。

- Originated from the Chinese script;
- Each kanji has at least 2 pronunciations – kun-yomi (Japanese pronunciation) and on-yomi (Chinese pronunciation);
- Provides semantics to words.

Different Tools for a Different Context

- We cannot simply tokenize based on spaces, because word boundaries are not that clear!

Different Tools for a Different Context

- We cannot simply tokenize based on spaces, because word boundaries are not that clear!
- Lattice-based tokenization is needed.

Lattice-based Tokenization

- Each connection has an associated cost;
- The most probable tokenization is the route with the lowest total cost.

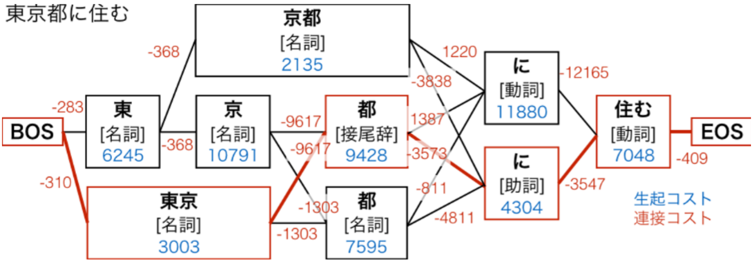


Figure: A lattice of tokens (Credit: Wanasit Tanakitrungruang)

Our Chosen Tokenizer

A widely-used dictionary for Japanese tokenization is MeCab, written in C++.

Our Chosen Tokenizer

A widely-used dictionary for Japanese tokenization is MeCab, written in C++.

Figure: Mekabu (Source: Wakasa no Himitsu)

Our Chosen Tokenizer

We shall use natto-py, self-described as “A Tasty Python Binding with MeCab”.

Our Chosen Tokenizer

We shall use natto-py, self-described as “A Tasty Python Binding with MeCab”.

Figure: Natto on rice. (Source: Wikipedia)

Aspect Detection

- We use natto-py together with CountVectorizer to extract out a list of the most common words in the entire dataset;
- Then assign each sentence in a review its relevant aspects.

Example output for natto-py

A token from natto-py looks like:

良かっ, 形容詞, 非自立,*,*, 形容詞アウオ段, 連用タ接続, 良い,
ヨカッ, ヨカッ

Example output for natto-py

A token from natto-py looks like:

良かっ, 形容詞, 非自立,*,*, 形容詞アウオ段, 連用タ接続, 良い, ヨカッ, ヨカッ

In English, a similar tokenization would look something like:

running, verb,*,*,*,*,present progressive tense,run,run,run

Aspect Keywords I

Aspect	Definition	Keywords
Service	Acts of help towards customer satisfaction	サービス, スタッフ, フロント, チェックイン, 丁寧, 親切, 接客, サーバ
Location	Access and landscape around the hotel	立地, 駅, バス, 近く, 便利, 駐車, コンビニ, 場所
Room	The attributes of the room	部屋, 広い, 宿泊, ベッド, 値段

Aspect Keywords II

Aspect	Definition	Keywords
Amenities	Other facilities in the hotel excluding room and bath	アメニティ, 無料
Bathroom	Bath in the hotel room, or a public bath	風呂, 温泉, 浴場, 露天風呂, 清潔, 湯, トイレ
Food	Morning or evening meals	朝食, 食事, 料理, 夕食, バイキング, メニュー, ご飯, 酒, 飲

Example: Aspect Identification in a Review

Review: “日本酒の飲み比べサーバは良かったです。また、部屋もきれいでスマホ充電など細かな気遣いも良かったです”

Example: Aspect Identification in a Review

Review: “日本酒の飲み比べサーバは良かったです。また、部屋もきれいでスマホ充電など細かな気遣いも良かったです”

Translation: “The server for the Japanese sake tasting was good. Likewise, the room was clean, and it was good that there was little to worry about things like smartphone chargers, etc. as well”

Example: Aspect Identification in a Review

Review: “日本酒の飲み比べサーバは良かったです。また、部屋もきれいでスマホ充電など細かな気遣いも良かったです”

Translation: “The server for the Japanese sake tasting was good. Likewise, the room was clean, and it was good that there was little to worry about things like smartphone chargers, etc. as well”

Aspects:

Service: “日本酒の飲み比べサーバは良かったです”

Food: “日本酒の飲み比べサーバは良かったです”

Room: “また、部屋もきれいでスマホ充電など細かな気遣いも良かったです”

oseti – A Japanese Sentiment Analysis Tool

- Dictionary-based – relevant words are matched with their sentiment polarity in a dictionary (positive, negative);

oseti – A Japanese Sentiment Analysis Tool

- Dictionary-based – relevant words are matched with their sentiment polarity in a dictionary (positive, negative);
- Text is analyzed at the sentence level; the overall sentiment is a simple weighted sum of positive and negative token sentiments;

oseti – A Japanese Sentiment Analysis Tool

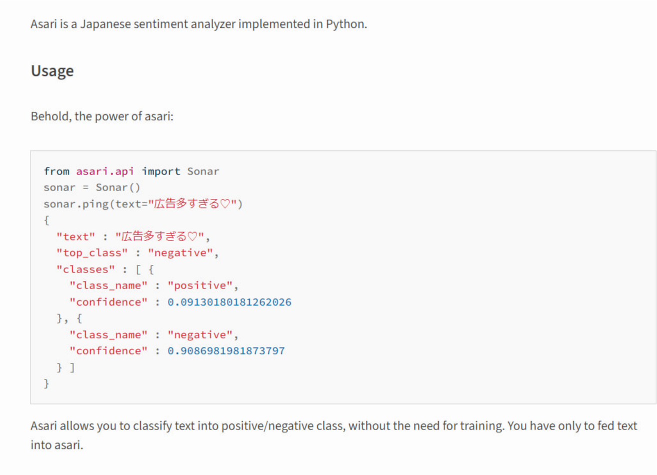
- Dictionary-based – relevant words are matched with their sentiment polarity in a dictionary (positive, negative);
- Text is analyzed at the sentence level; the overall sentiment is a simple weighted sum of positive and negative token sentiments;
- Does not consider position of words in a sentence.

oseti – A Japanese Sentiment Analysis Tool

- Dictionary-based – relevant words are matched with their sentiment polarity in a dictionary (positive, negative);
- Text is analyzed at the sentence level; the overall sentiment is a simple weighted sum of positive and negative token sentiments;
- Does not consider position of words in a sentence.

Figure: A Japanese *oseti* (Source: Wikipedia)

asari – Another Japanese Sentiment Analysis Tool



Asari is a Japanese sentiment analyzer implemented in Python.

Usage

Behold, the power of asari:

```
from asari.api import Sonar
sonar = Sonar()
sonar.ping(text="広告多すぎる♡")
{
  "text" : "広告多すぎる♡",
  "top_class" : "negative",
  "classes" : [ {
    "class_name" : "positive",
    "confidence" : 0.09130180181262026
  }, {
    "class_name" : "negative",
    "confidence" : 0.9086981981873797
  } ]
}
```

Asari allows you to classify text into positive/negative class, without the need for training. You have only to fed text into asari.

Figure: Documentation for asari

asari – Another Japanese Sentiment Analysis Tool

- Trained using a Linear Support Vector Classifier;

asari – Another Japanese Sentiment Analysis Tool

- Trained using a Linear Support Vector Classifier;
- Returns the positive and negative sentiments as probabilities (confidence levels).

asari – Another Japanese Sentiment Analysis Tool

- Trained using a Linear Support Vector Classifier;
- Returns the positive and negative sentiments as probabilities (confidence levels).

Figure: Asari clams with udon (Source: Wikipedia)

Sentiment Analysis Methodology

1. We take each of the aspects contained in the review, and assign two sentiment scores using oseti and asari respectively (per aspect);

Sentiment Analysis Methodology

1. We take each of the aspects contained in the review, and assign two sentiment scores using oseti and asari respectively (per aspect);
2. Convert each sentiment score (a number from -1 to 1) into a rating from 1 to 5;

Sentiment Analysis Methodology

1. We take each of the aspects contained in the review, and assign two sentiment scores using oseti and asari respectively (per aspect);
2. Convert each sentiment score (a number from -1 to 1) into a rating from 1 to 5;
3. The ratings given by the customer for their reviews will be taken as a “ground truth”, and the predicted scores will be compared against them.

Example Analysis of Review

Let's look at the "food" aspect of the following review:

家族でのんびり過ごせて良かったです。お料理が美味しかったし、景色も美しかったです。温泉は思ったよりも小さかったですが、泉質は結構よかった。

Translated, it reads:

I had a relaxing stay with my family. The food was delicious, and the scenery was beautiful too. Though the hot spring was smaller than I thought, the spring water was rather good.

Example Analysis of Review

Let's look at the "food" aspect of the following review:

家族でのんびり過ごせて良かったです。お料理が美味しかったし、景色も美しかったです。温泉は思ったよりも小さかったですが、泉質は結構よかった。

Translated, it reads:

I had a relaxing stay with my family. The food was delicious, and the scenery was beautiful too. Though the hot spring was smaller than I thought, the spring water was rather good.

Analyzer	Sentiment	Predicted Score
oseti	1	5
asari	0.887214	5

Model Evaluation

As classifiers, neither sentiment analyzer performs well. However, a prediction of 4 stars when the actual customer rates 5 stars is still not too bad.

Model Evaluation

As classifiers, neither sentiment analyzer performs well.
However, a prediction of 4 stars when the actual customer rates 5 stars is still not too bad.
Thus, we use the *mean absolute error* as a metric to gauge the performance of both models.

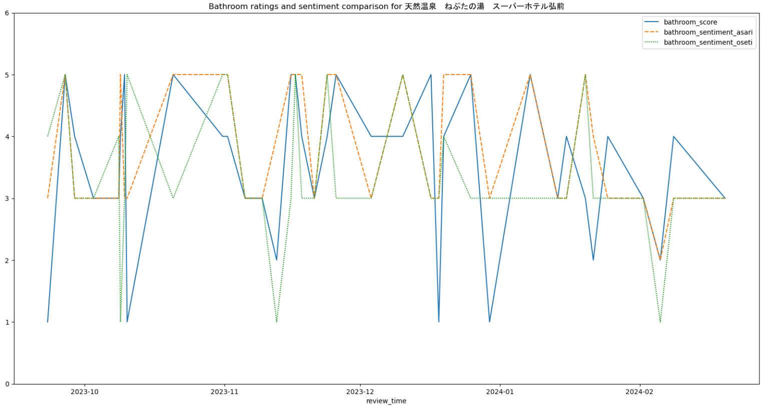


Figure: Plot of sentiments and ratings

Aspect-Based Comparison of Mean Absolute Errors

	oseti	asari
Overall	0.836027	0.600395
Service	1.054956	1.008441
Location	1.183908	1.095725
Room	1.072557	0.953125
Amenities	1.086745	1.084590
Bathroom	0.943426	0.839260
Food	0.947198	0.795617

Limitations

- More time, more data;

Limitations

- More time, more data;
- Proper labelling of aspects;

Limitations

- More time, more data;
- Proper labelling of aspects;
- Accounting for zero anaphora; 選んでいただいたものなら何でも結構です。(*erande itadaita mono nara nandemo kekkou desu*, lit. “Whatever (you) pick is fine.”) – aspects may not appear even though the customer intended it.

Future Work

- Collecting data from all prefectures of Japan, perhaps from other countries as well;

Future Work

- Collecting data from all prefectures of Japan, perhaps from other countries as well;
- Utilize a neural network approach to find hidden meanings;

Future Work

- Collecting data from all prefectures of Japan, perhaps from other countries as well;
- Utilize a neural network approach to find hidden meanings;
- Work out a translation scheme that keeps the intent as accurately as possible.

Summary

1. Introduction
2. Data Processing
3. Modelling
4. Limitations and Future Work

The End

Thank you.