Natural Language Project
David Blankenship

Pitch:

The goal of my project is to create, demonstrate, and provide a proof of concept of a pipeline for Optical Character Recognition and Machine Translation in Japanese. The target audience would be for publishers looking to lower costs in translating foreign language print media and, hopefully, allow otherwise previously financially unviable media to be brought to new markets. Similarly, there is a hope that the ability to reach wider audiences can be given directly to the hands of the original creators themselves.

Data Source:

I sourced random pages of Japanese language manga from twitter. I will include them here in my report but will not be able to post them in my code to be posted to github as, while they were posted on twitter, they are copyrighted works, and I don't have permission to share them widely. I will note the image filename that should allow you to figure out which portions of my code correspond to what.

Model and Data Justification:

MangaOCR was selected for its ease of use and optimization for use in manga, which is the primary focus for this project. It also consistently found the entire sentence at least when the image was clipped to only include the text inside speech bubbles. I give a link to its github page below

https://github.com/kha-white/manga-ocr

I also tried EasyOCR but found that it would only identify a few characters in a text bubble and fail to capture the whole sentence.

https://github.com/JaidedAI/EasyOCR

For the translation models I tried using 4 translators all available at huggingface:

- Helsinki-NLP/opus-mt-ja-en
- staka/fugumt-ja-en
- facebook/nllb-200-distilled-600M
- facebook/nllb-200-1.3B

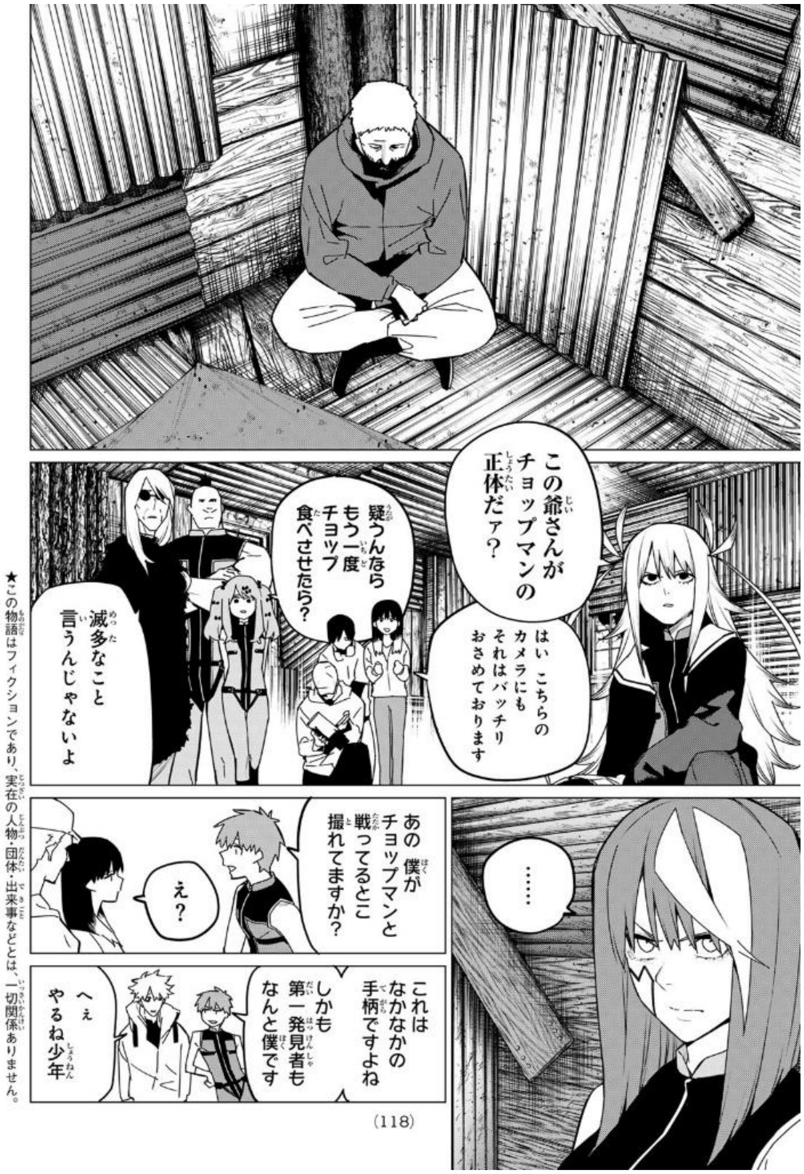
I will not reproduce it here, but you will see in my code that only the Helsinki-NLP model gave consistently good results. The others all gave what can only be described asgarbage. In fact, for some bizarre reason the facebook nllb models returned results in Spanish, English, and Japanese. I'm frankly at a loss of words for this behavior as I specified it was a ja to en translation. There even appears to be what looks like artifact of the language it's supposed to be translated to in it so I suspect there is a bug of some sort.

Natural Language Project
David Blankenship

Commented Examples:

I include the original image and, where I have it the human translations.

1st example: random_manga1


この爺さんがチョップマンの正体だァ？
はい こちらのカメラにもそれはバッチリおさめております
疑うんならもう一度チョップ食べさせたら？
滅多なこと言うんじゃないよ
★この物語はフィクションであり、実在の人物・団体・出来事などとは、一切関係ありません。
……
あの 僕がチョップマンと戦ってるとこ撮れてますか？
え？
これはなかなかの手柄ですよね
しかも第一発見者もなんと僕です
へぇ やるね少年
(118)

To start with the analysis, we look at the output of the OCR. It should be noted that using the full page all at once gives very poor results. I instead clipped out the text boxes where it makes sense to. Also, not all sound effects are included in these examples.

'この爺さんがチョップマンの正体だァ？'

'はいこちらのカメラにもそれはバッチリおさめております'

'疑うんならもう一度チョップ食べさせたら？'

'滅多なこと言うんじゃないよ'

'あの僕がチョップマンと戦ってるとこ撮れてますか？'

'え？'

'これはなかなかの手柄ですよね'

'しかも第一発見者もなんと僕です'

'へぇやるね少年'

You can review yourself, but you will see that the OCR was excellent in recognizing the text.

The machine translation from our pipeline was as follows:

"This old man is the identity of the chopman, isn't he?"

'(Laughter)'

"If you don't believe me, why don't you let him eat his chop again?"

"Don't say that often."

'Are you taking pictures of me fighting the chopman?'

'What?'

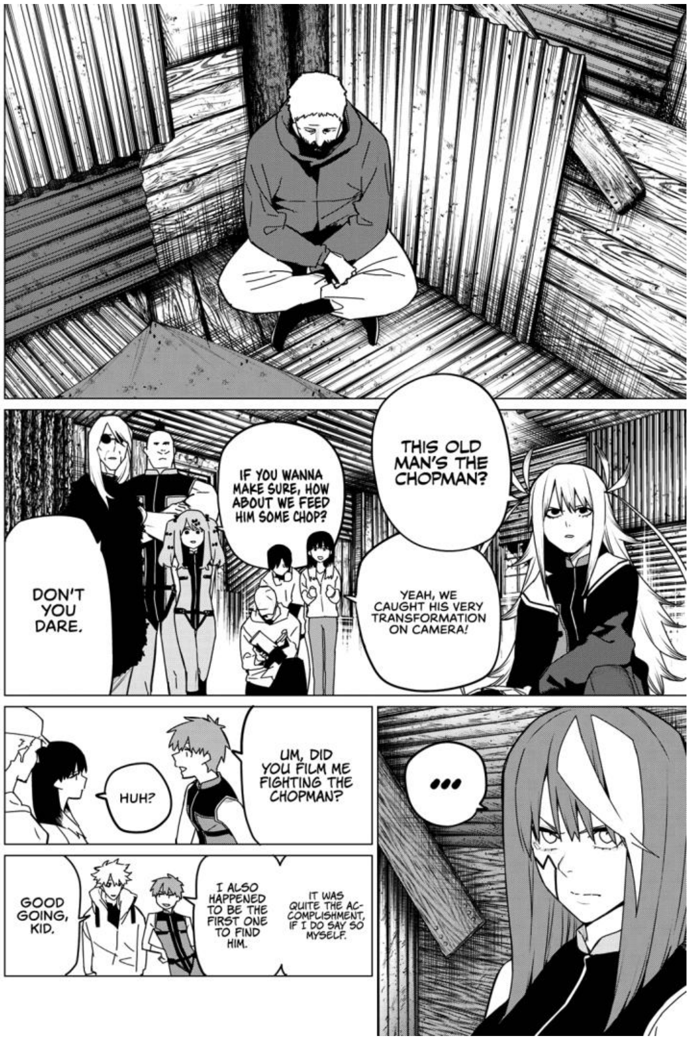
"It's quite an honor, isn't it?"

'(Laughter)'

'Well done, boy.'

Obviously, there are problems. (Laughter) appears to be a failure state for translations. While initially testing for not having to cut up everything in the second example it found nonsense words which were also translated to (laughter). However it’s not all bad!

For instance, the phrase 'この爺さんがチョップマンの正体だァ？’ translated to "This old man is the identity of the chopman, isn't he?" On the surface this is nonsense, but the katakana チョップマン actually do sound out Chopman (or at least how the Japanese would pronounce it). Katakana are typically used for western loanwords so this actually checks out, they actually said something about a guy called Chopman. And sure enough looking at the image the characters are looking at some dude, presumably Chopman himself. We compare with the translation from a fan translator below and see they chose “This old man’s the Chopman?”. We can see what the machine translation is getting at!


THIS OLD MAN'S THE CHOPMAN?
YEAH, WE CAUGHT HIS VERY TRANSFORMATION ON CAMERA!
IF YOU WANNA MAKE SURE, HOW ABOUT WE FEED HIM SOME CHOP?
DON'T YOU DARE.
...
UM, DID YOU FILM ME FIGHTING THE CHOPMAN?
HUH?
IT WAS QUITE THE AC-COMPLISHMENT, IF I DO SAY SO MYSELF.
I ALSO HAPPENED TO BE THE FIRST ONE TO FIND HIM.
GOOD GOING, KID.

Using the above to compare the BLEU score from huggingface we found the following results:

{'bleu': 0.0, 'precisions': [0.26, 0.0975609756097561, 0.02857142857142857, 0.0], 'brevity_penalty': 0.8025187979624785, 'length_ratio': 0.819672131147541, 'translation_length': 50, 'reference_length': 61}

Which is quite bad. And I am torn on this one. Some of the machine translations seem to be decent enough, or at least basically accurate, even if some of them are quite bad.

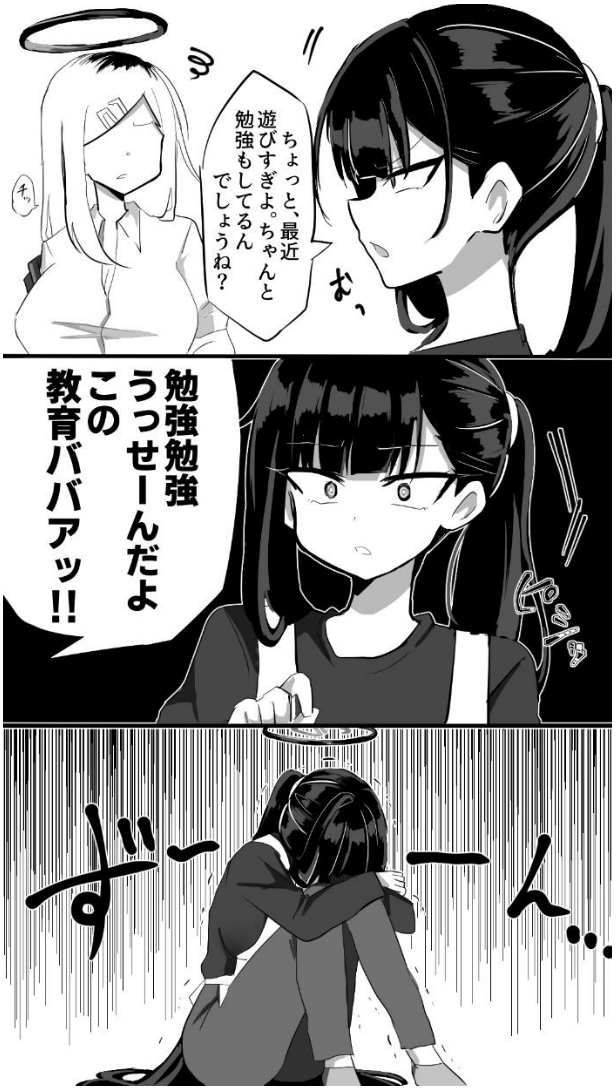
For example, “Good going, kid” and “Well done, boy” are basically the same and after a bit more investigation “Well done, boy” is more correct. “少年” pronounced shounen means “boy” in Japanese and is precisely the word used in the Japanese text. In the next section and an addendum, I add additional context about the Japanese to English translation community about matters like this as they are affecting the ability to score these.

I will finally note that I also attempted to concatenate all of the sentences from the OCR prior to translation but found little success with this method. See below what I received and yes the repeated i&gt; was part of what I got back.

Natural Language Project
David Blankenship

"This old man is the identity of the chopman, and he's got it in his camera, and if you're suspicious, you shouldn't tell him that I'm fighting the chopman. &lt;i&gt;That's a pretty good shot.&lt;/i&gt; &lt;i&gt;That's pretty impressive.&lt;/i&gt; &lt;i&gt;That's great.&lt;/i&gt; &lt;i&gt;That's great. &lt;i&gt;That's great. &lt;i&gt;That's great. &lt;i&gt;i&gt;That's great. &lt;i&gt;i&gt;That's great. &lt;i&gt;i&gt;i&gt;that's great.i&gt;i&i&gt&gt&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&gt;i&i&i&gt;i&gt"

Second Example: twitter_manga1


ちょっと、最近遊びすぎよ。ちゃんと勉強もしてるんでしょうね？
む、
チッ
勉強勉強うっせーんだよこの教育ババアッ!!
ず〜ーん…

I first broke this up into the three panels and received the following:

'それでも、'

'そういうことで、今回の勉強勉強うっせーんだよ教育ババアッ！！'

'ずーん．．．'

Translated as:

'(Laughter)'

"That's why you're studying so hard this time!"

"It's just..."

The first index isn't capturing anything. Although I suspect it may be capturing a visual emotion indicator at the top left corner of the image as the そ hiragana looks somewhat similar. It also appears to again be translating it as laughter which is again obviously wrong.

The second index starts poorly and doesn't seem to actually be capturing anything. However, starting at the 勉 kanji it briefly gets it right in the second panel, misses a short two hiragana line, and then gets the remainder correct.

The third line is impressively correct as it is in a nonstandard format on the page. I actually expected it to fail.

After this I cut into the text boxes as in the first. This simply seems to be the best way to handle it for our OCR used. I received the following:

'ちょっと、最近遊びすぎよ。ちゃんと勉強もしてるんでしょうね？'

'チッ'

'勉強勉強うっせーんだよこの教育ババアッ！！'

'ずーん．．．'

Translated as:

"You've been playing too much lately. You're also studying well, aren't you?"

'Shit.'

"I'm going to study! I'm going to study! I'm going to study!"

"It's just..."

Most of the Japanese text matches up! Possibly even all the second character in 'チッ' may be the correct one, but the font in the actual image makes it very hard to tell.

Natural Language Project
David Blankenship

I did not have a human translated page for this so I elected to try using Google translate and Chat GPT. For Google translate I received:

- Hey, I've been playing too much lately. You're also studying properly, aren't you?
- Chit
- Study, study, study, study,
- Hmmm....

From Chat GPT I provide two translations All at once, and line by line. This is when I ask it to translate all at once:

- "Hey, you've been playing around too much lately. You're studying properly too, right?"
- "Tch."
- "You're always nagging about studying, you education hag!!"
- "Thud..." (sound of being emotionally hurt)

When I ask it to translate line by line, I get the below. I also include notes on each line.

- "Hey, you've been playing too much lately. You're studying properly too, right?"
- "Tsk" or "Tch," which is a sound indicating annoyance or irritation.
    - The translation as “shit” is arguably good as its basically a rude way to express annoyance. It may even be just mistaking an onomatopoeia as チッ sounds like chitsu. Which may either be an onomatopoeia for tongue clicking or the actual English loan word. Both would be expressed in katakana as this is!
- "Studying, studying, you're so annoying, you old hag!!"
    - We can see that the machine translation getting stuck on saying "Studying" is probably because it is actually repeated. See 勉強勉強 which are the same two kanji, probably meaning studying, repeated twice.
- "Sigh..." or "Thud..." indicating a feeling of heaviness, depression, or disappointment.
    - From this I think shit and even chit are okay translations, conveying the idea of annoyance or possibly even the literal onomatopoeia respectively.

Also, the ability of ChatGPT to translate to a very readable format is quite impressive. I'll note as well that I had forgot to ask it to translate when I first entered it and found it responding to me in Japanese.

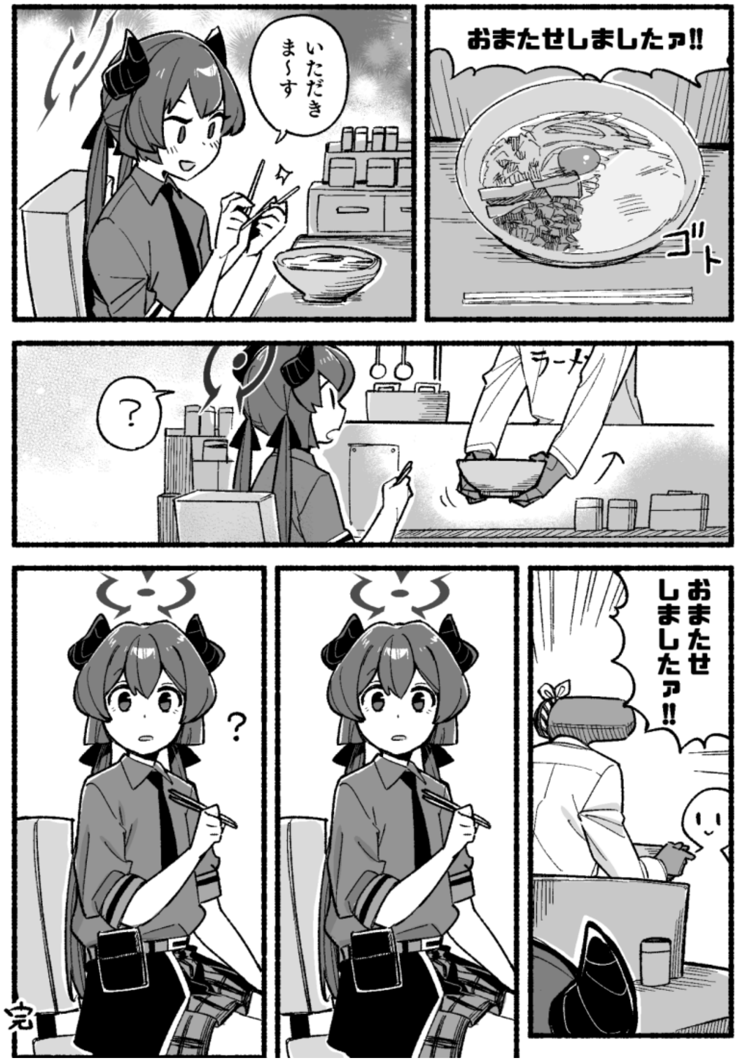
Regardless I used my judgement for the various translations and choose the chat gpt line by line as the most reasonable for use in BLEU.

```
{'bleu': 0.2081918849957486, 'precisions': [0.25925925925925924, 0.21739130434782608, 0.2,
0.16666666666666666], 'brevity_penalty': 1.0, 'length_ratio': 1.2272727272727273,
'translation_length': 27, 'reference_length': 22}
```

We see now that we are getting an above 0 score!

Third Example: Junko manga

We now look at a third page which I, quite ironically, chose because I thought it would be very straightforward. In addition to only having only three word bubbles, two of them are just repeated for the joke. As we see here, in the testing section, and an addendum this page resulted in me writing quite a bit!



The OCR found the following:

'おまたせしましたァ！！'

'いただきま～す'

'おまたせしましたァ！！'

Which is entirely correct.

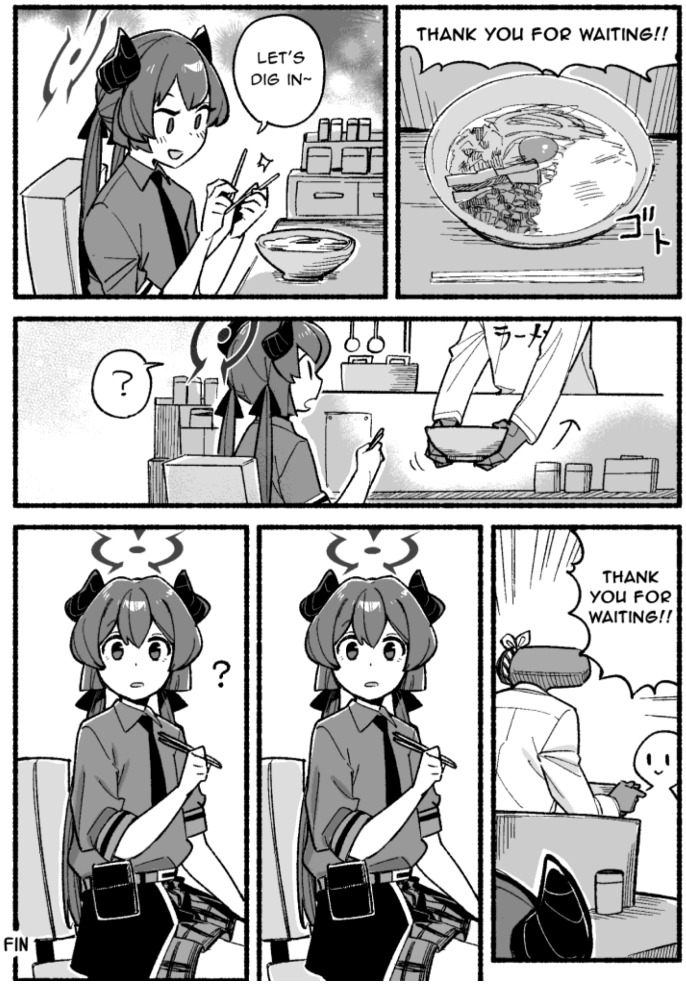
Natural Language Project
David Blankenship

The machine translation gave:

"I'm sorry for letting you go!"

'Thank you for the food.'

"I'm sorry for letting you go!"



I have an extensive discussion in the addendum on my issues with the human translation and my thoughts on the accuracy of the machine translation. I will not repeat it here. Instead, I show the results of score where my opinion on the translation is taken into account followed by the human translation solely:

```
{'bleu': 0.31867539330789935, 'precisions': [0.5294117647058824, 0.2857142857142857, 0.2727272727272727, 0.25], 'brevity_penalty': 1.0, 'length_ratio': 1.5454545454545454, 'translation_length': 17, 'reference_length': 11}
```

```
{'bleu': 0.0, 'precisions': [0.23529411764705882, 0.0, 0.0, 0.0], 'brevity_penalty': 1.0, 'length_ratio': 1.5454545454545454, 'translation_length': 17, 'reference_length': 11}
```

Testing:

I ran into a serious issue when using BLEU due to its need for a reference translation and I wrote an almost 1000 word essay in my Jupyter Notebook explaining my issue with localization and the culture of translators of Japanese to English in the West that is clearly having an effect on my ability to score the translations accurately. I reproduce the full argument below in an addendum but will summarize here.

There is a culture amongst professional (and even a subset of hobbyist) western-based translators of Japanese to strip cultural context from their translations and arbitrarily edit or rewrite sentences to better fit their own preconceived notions. This makes human translations far less useful for judgements of quality than they should be.

I demonstrate the issue in the 3rd example where the phrase “いただきます” or “itadakimasu” is, in my opinion, mistranslated by the human translator as “lets dig in” rather than “Thank you for the food”. I suggest the latter phrase is a far superior translation and was precisely what the machine translation provided. Due to this localization, as this tendency is known, we now receive a worse score for a better translation and will struggle to find quality translations for developing machine translation technology for Japanese to English.

I also discovered while writing the Examples section (which I wrote after this originally) another issue of scoring using BLEU in the 1st example where “kid” is used by a human translator whereas “boy” is the literal word used. I give a summary of all the scores I got but feel ever less confident that I can reasonably expect these to be able to reflect accurately the quality of translation due to human-side issues which I find frankly astounding.

I mention these arbitrary changes in my discussion in the addendum, but this project has opened my eyes to the full scale of arbitrariness of these changes in the translations. I had figured these comics would be straightforward to translate but am shocked to discover the issues here.

1st example: random_manga1

```
{'bleu': 0.0, 'precisions': [0.26, 0.0975609756097561, 0.02857142857142857, 0.0],
'brevity_penalty': 0.8025187979624785, 'length_ratio': 0.819672131147541, 'translation_length':
50, 'reference_length': 61}
```

2nd example: twitter_manga1

```
{'bleu': 0.2081918849957486, 'precisions': [0.25925925925925924, 0.21739130434782608, 0.2,
0.16666666666666666], 'brevity_penalty': 1.0, 'length_ratio': 1.2272727272727273,
'translation_length': 27, 'reference_length': 22}
```

3rd example: Junko manga

```
{'bleu': 0.31867539330789935, 'precisions': [0.5294117647058824, 0.2857142857142857,
0.2727272727272727, 0.25], 'brevity_penalty': 1.0, 'length_ratio': 1.5454545454545454,
'translation_length': 17, 'reference_length': 11}
```

```
{'bleu': 0.0, 'precisions': [0.23529411764705882, 0.0, 0.0, 0.0], 'brevity_penalty': 1.0, 'length_ratio':
1.5454545454545454, 'translation_length': 17, 'reference_length': 11}
```

First, precision for 1-gram tokens is consistently the highest across all translations and it consistently decreases from there as the number of tokens looked at increases. I wonder if this is from the nature of the Japanese language in which a single token can have a fair bit of meaning or if this is typical for all languages.

For my BLEU scores none of them are particularly high but given the human-side issues I have discussed and some of the frankly very clear issues with the machine translator this is not surprising.

Code and instructions to run it:

When looking at my code, it helps to have some understanding of the Japanese writing system. I include a short summary in my Jupyter Notebook and will be included it in an addendum below.

My code is linked here:

https://github.com/General-Cow/MangaAutoTranslator

In order to run my code, there are three major functions; get_translations(), reference_wrapper(), and get_evaluations(). Get_translations() is made such that either an individual image or directory of images can be pointed to and then read in by the OCR and translated automatically. Reference_wrapper is for use inside of get_evaluations() to wrap the reference translations for BLEU into the frankly annoying format required. This allows the user to write the reference list in a much less annoying, much more intuitive way. Get_evaluations takes in the machine translation from get_translations() and a reference list and outputs the bleu evaluation.

To actually run my code, simply use the get_translations() function, create a reference list of acceptable translations, followed by the get_evaluations() function. Nothing more is required.

It should be noted that my code defaults to the Helsinki-NLP/opus-mt-ja-en model for Japanese to English translations but does not require it. In fact, my entire pipeline is language and model agnostic! Feel free to use it in any other language you prefer!

Addendum: Japanese Writing System

It is useful to describe the Japanese writing system briefly to understand the analysis I have in this project. There are three major elements for writing Japanese; Hiragana, katakana, and kanji. The first two are phonetic syllabaries while kanji is a logographic system derived from, but distinct to, Chinese characters where individual kanji represent specific words or concepts. Hiragana is typically used for native Japanese words while katakana is used for loanwords and various special uses. Kanji may sometimes be accompanied by small text next to it. These are called furigana, and they are a reading aid used to show the pronunciation of kanji, usually for less frequently used kanji or in works expected to be read by children still learning their kanji. All three of these can be used in a sentence together. Kanji can also be used as part of a word with hiragana to form a word, for example 休む is a single word and means 'to rest' and is a kanji (休) and a hiragana character (む).

It should also be noted that there are no spacings between words or sentences. They use similar punctuation although some such as their periods, commas, or quotations look slightly different. Some such as question marks are loaned from the West but are only used in informal writings.

The order for reading Japanese is also important. Japanese can be written vertically or horizontally. If it is written vertically then the sentence starts from the top, goes to the bottom and is ordered from right to left. This is the traditional style of writing and is called tategaki. Japanese can also be written horizontally and is read from left to right as in English. This is the modern style and is called yokogaki.

We will typically see images written in tategaki, which is the traditional format for manga, while the output of our OCRs will be in yokogaki, the normal style for Japanese writing on computers. These are not hard rules however!

More can be read here: https://en.wikipedia.org/wiki/Japanese_writing_system

Addendum: Localization

Note: This is in relation to the translation in the third example of "いただきます" or "itadakimasu".

A brief note: the 〜 symbol in the original image and OCR text is just a symbol that shows the preceding character was drawn out in it's pronunciation. Think "Hello" as opposed to "Helloooo" in English.

The reproduced text follows:

So the human translation translated it as:

"Thank you for waiting!!"

"Let's dig in~"

Natural Language Project
David Blankenship

"Thank you for waiting!!"

They are taking some liberties with the translation. The second line is a phrase that the Japanese say before eating a meal that is more accurately a sort of prayer of gratitude for the food. It's very often translated as "Let's dig in" but I would suggest that this is a mistranslation. "いただきます" or "itadakimasu" (the phonetic pronunciation of the phrase in question) literally means "to humbly receive" and is a prayer coming from the Buddhist influence on Japanese culture and is often said with hands clasped together and eyes closed, not unlike a Christian prayer before a meal. "Thank you for the food" is almost certainly a more accurate translation than the flippant "lets dig in".

Something I've avoided talking about at this point but now must is that there's a very long running argument over localization in Japanese translations particularly in manga. Being charitable, localization is where, rather than a literal translation, the work is slightly edited or rewritten so that it can be understood better by its target audience. For example, Japanese idioms or puns won't make any sense without a translators note. Less charitably, it erases the unique Japanese character and authorial intent of these works in an attempt to appeal to wider audiences and/or reflect the translators personal (or frankly sometimes political) disdain of certain aspects of Japanese culture that differ from ours. In my personal opinion (and very many others), localization can go so far as to completely alter the meaning of a work and ignores the original appeal that drew people to them in the first place.

I bring this up as I tried a Chat GPT input for translating this phrase and received "Let's eat!" or "Bon appétit!". I then asked, "what does itadakimasu mean? and what is the cultural context" and received the following two paragraphs:

> "Itadakimasu" (いただきます) literally means "I humbly receive." In the cultural context of Japan, it is a phrase traditionally said before eating a meal. It conveys gratitude and appreciation for the food, the people who prepared it, and the ingredients, which may have involved the sacrifice of animals or plants.
>
> The phrase is deeply rooted in Japanese etiquette and respect for the effort and life involved in providing the meal. It reflects a broader cultural emphasis on mindfulness and appreciation. Saying "itadakimasu" before a meal is a way to express humility and thankfulness, aligning with the Japanese values of respect and gratitude.

Does this seem like a phrase that "lets dig in" has accurately conveyed? No, it does not. Only in the most tortured form of "a phrase you can say before you eat" where any and every last ounce of cultural context has been stripped away completely.

This suggests to me that localization in translation is possibly altering the results for when you ask ChatGPT to translate something that strips cultural context. These localized sources could be what it's pulling from when you ask it specifically to "translate". From my description above I obviously have a personal preference towards this issue, and one could reasonably criticize me for overreacting in this specific instance, however I have seen significantly mistranslated works sometime even to the point of being precisely the opposite of the original. I would suggest that if there is a large body of mistranslations being used to teach machine translation algorithms or LLMs that is a very real problem; garbage in, garbage out as they say. I still believe that ChatGPT has potential has potential as part of a pipeline, but it is concerning to see this issue.

I should note that ChatGPT translated the first and last sentence as "Sorry to keep you waiting!!" which is probably fine. As is the human translation for these lines. Probably. Unfortunately, as I again lack the ability to speak or read Japanese fluently, I can't say which is better though I suspect from context that the chef in this image is using a specific Japanese idiom/phrase that may literally translate to something including an apology for waiting but can reasonably be understood to be similar to thanks for waiting. This would be an excellent example of localization done right if what I suspect is correct. Although on the other hand, what precisely is lost if you keep it as "Sorry to keep you waiting?" which is possibly a more literal translation.

I will add this issue of localization is making scoring these using BLEU very frustrating. How can I reasonably score Japanese translations when there seems to be a culture amongst Western-based translators of Japanese that they retain a right to seriously alter the meaning of words and sentences in their translation according to their arbitrary whims. I'm not aware of any language other than Japanese that receives this treatment and it's not for lack of looking or, as I'm sure is obvious, passion about this subject. "Let’s dig in" is a mistranslation, full stop. Yet the machine translator I used correctly gave me "Thanks for the food". If I naively use this, I am going to get a worse score for a correct translation because someone wanted to wash away the religious and cultural context of the phrase "itadakimasu".

I will even test this to illustrate my point.

This is followed by me showing that if we include only the “lets dig in” mistranslation that indeed the score does fall despite our machine translation being perfectly valid and frankly better. See example 3 above. I finished with an additional final word below:

And exactly as I said, the poor human translation is now causing the score to be worsened. We have moved beyond a mere simple preference for one translation style over another. Because of this practice, we may find that developing machine translation technology to translate Japanese text is now harder than it should be since they now have to contend with the poor practices of professional translators.