Evaluation of Language Models in Multilingual Machine Translation

1 Introduction

In this final project, we use several language models to perform machine translation between several languages. We cover four pre-trained language models and discuss their architectures and implementation, compare their performances and analyze some of the errors in outputs.

MMT is a model that translates text from many source languages to many target languages. MMT is a nontrivial task as languages are vastly diverse, and there are many language pairs as opposed to one. The inconsistent availability of data makes MMT a more complex problem to solve than translation between just two languages. Some languages have abundant parallel training data, while others have little data and are considered low-resource languages. There are cultural references in all languages which are difficult to translate, as not all languages may have the same ideas. Additionally, the word order (such as subject-verb-object) varies among languages. MMT is an important problem to solve, as it will make machine translation more efficient and accessible.

With a well-trained MMT model, instead of training separate models for each language pair, it is possible to utilize one model to translate many different pairs of languages. This is very memory- and time-efficient compared to training several bilingual machine translation models, which makes MMT an increasingly interesting topic.

2 Related work

(Zhu et al., 2023) studied several different LLMs on multilingual tasks, and discovered that ChatGPT outperforms other LLMs, but still lags behind state-of-the-art bilingual models. This performance gap is mostly attributed to poor performance on low-resource langauges and translating English into non-English languages. Similarly, (Aharoni et al., 2019) evaluated the effect of training transformer models to translate over 200 language pairs, and found that models trained only on a few languages performed better on average than models trained on a greater number of languages. Still, they concluded that, even when dealing with such a large amount of languages, the models were able to work reasonably well.

(Dabre et al., 2020) describes various approaches to training multilingual machine translation models and also explores their benefits over bilingual models. As was the case in the previously mentioned papers, they found that, with existing models and methods, performance is reduced when trying to train a single model to translate multiple different languages. However, multilingual models also have some benefits to them: multilingual models are helpful for translating low-resource languages, as they are able to draw on their knowledge of other languages to help. Furthermore, multilingual models are more compact than using a multitude of bilingual models, leading to less computational costs.

Similar to our approach, (Vilar et al., 2022) studied the performance of PALM (Chowdhery et al., 2022) on WMT data. However they focused on only a few select languages such as French, German and Chinese. Instead of analyzing the WMT dataset as-is, they look at more recent WMT21 data and also evaluate on more refined subsets. They discovered that PALM seems to be better at translating into English than the other direction. Also while PALM seems to be good at creating fluent sentences, it is less accurate than state-of-the-art machine translation models.

(Hendy et al., 2023) performed a similar evaluation on the performance of ChatGPT in machine translation. They compared the performance of ChatGPT on the WMT22 dataset against the top scoring models for each language pair. They found that ChatGPT was able to produce fluent translation outputs even in the zero-shot setting for high-resource languages, and found that they could improve performance by taking advantage of its in-context learning ability by providing a few labelled examples along with the test input.

One noteworthy phenomena in multilingual machine translation is negative interference, where training models on concatenated multilingual data results in lower performance on high-resource languages. (Wang et al., 2020) show that negative interference could also happen for low-resource languages. They suggest that the negative interference could be caused by parameters related for different languages competing for capacity.

3 Dataset used

We use the WMT2014 Translation dataset, which consists of 5 language pairs(French-English, Hindi-English, German-English, Czech-English, Russian-English). The data is mostly news stories taken from the Europarl(European Parliament Proceedings Parallel Corpus) corpus, the News Commentary Corpus, and the Common Crawl Corpus. For Hindi, data is crawled from HindEnCorp. We utilize the test data split(2507 sentences for the Hindi-English pair, 3003 sentences for other language pairs) for analysis of our pre-trained language models. This is a fairly non-trivial task as we evaluate on all ten directions. Below are a few examples of the dataset:

Table 1: Examples of WMT14 dataset entries

Language 1	Language 2	Content 1	Content 2
English	Czech	A black box in your car?	Černá skříňka ve vašem autě?
English	French	The tea party is aghast.	Le Tea Party est atterré.
English	German	It was not something people wanted.	Die Leute wollten es nicht.

4 Baselines

We use OPUS-MT as our baseline. OPUS-MT is a collection of pretrained neural machine translation (NMT) models, based on Transformers and deep RNNs. More detail on Opus-MT can be found in the Approach section.

5 Approach

We evaluated four models in our analysis: GPT-3.5, PALM2, mBART-50 and Opus MT on the WMT14 dataset. GPT-3.5 and PALM2 are both LLM(Large Language Model)s that are not directly available, and are only available for inference through API requests. Opus MT is an open-source collection of pre-trained language models based on transformers. They are specifically trained for translation, and can be directly used to translate text. mBART-50 is a multilingual machine translation model based on the original BART that was developed at Facebook AI. It has been trained on 50 languages, and can also be used through the Transformers library.

5.1 Opus-MT

OPUS-MT (Tiedemann and Thottingal, 2020) is not a translation model itself, rather a set of translation tools and services. It is based on a neural translation framework called MarianNMT. MarianMT is an NMT model based on deep RNN and transformer that was developed at Microsoft, and it is the main engine behind Microsoft Translate services.

MarianNMT is based on the encoder-decoder architecture. It captures relationships between words in a sentence using self-attention mechanisms and positional encoding.

Opus-MT is simply the MarianNMT model pre-trained on Open Parallel Corpus (OPUS), which is a collection of human-translated texts from the Internet. It contains content from books, articles, movie subtitles, and others. Preprocessing is all done automatically without manual intervention.

To use Opus-MT, we use the pipeline capability from the Transformers library. This simplifies the process of tokenizing the input and generating the translation output into one step. Because of MarianNMT's lightweight implementation and high efficiency, Opus-MT was the only model we were able to run locally. It took about 10 hours to translate all pairs on a laptop.

5.2 GPT-3.5

We use OpenAI's API to gain translations from GPT-3.5, the same model used in ChatGPT. GPT-3.5 is a fine-tuned version of GPT-3(Brown et al., 2020), a huge Transformer-based model with 175 billion parameters. GPT-3 was primarily trained on Common Crawl data, out of which 93% of tokens were English.

There were several difficulties in utilizing the OpenAI API: with a free account only we could only send 3 RPM(requests per minute), which would result in each language pair(3003 sentences) taking over 16 hours to process assuming that the OpenAI server responded immediately to

Table 2: Performance (BLEU) with WMT14: full dataset

Model	de-en	fr-en	cs-en	hi-en	ru-en	en-de	en-fr	en-cs	en-hi	en-ru
GPT-3.5	**34.9**	**38.07**	**36.34**	**25.82**	**38.10**	**28.50**	**41.28**	25.84	**17.90**	**36.36**
Opus MT	33.93	37.98	32.59	13.40	31.91	27.56	39.85	**27.30**	9.94	32.59
mBART-50	32.61	36.78	34.55	25.61	36.46	24.93	35.34	24.07	17.44	32.05

Table 3: Performance (BLEU) with WMT14: subset

Model	en-de	en-fr	en-cs	en-hi	en-ru
PALM2	**28.84**	38.69	26.05	**23.26**	**36.46**
GPT-3.5	28.51	**41.31**	25.84	17.90	36.33
Opus MT	27.58	39.88	**27.33**	9.96	32.62
mBART-50	24.95	35.36	24.09	17.43	32.05

requests. After some experimentation this soon turned out to be unfeasible, so we set up a paid account which allowed for 3,500 RPM and used a Python script provided by OpenAI to send parallel requests[1]. This greatly reduced inference time for each language pair to about 1.5 hours. The model was prompted in the following format: *Translate the following sentence from* $\{$*source language*$\}$ *to* $\{$*target language*$\}$*:* ```$\{$*text to translate*$\}$```. Thus the model was provided with which language the input sentence was along with the target language to translate into. The triple backticks surrounding the source sentence were inserted as per advice from an online course discussing best practices when prompting ChatGPT [2].

As can be seen in Table 2, GPT-3.5 performs significantly worse for Hindi, and is significantly worse at translating other languages into English compared to vice-versa.

5.3 PALM2

PALM2(Anil et al., 2023) is the recently unveiled successor of PALM(Chowdhery et al., 2022). A Transformer(Vaswani et al., 2017)-based model trained on hundreds of languages and diverse domains, it collects data from a wide array of sources such as the web, books, and code. Notably it is trained on a higher proportion of non-English data compared to previous LLMs.

PALM2 is currently available through MakerSuite, and requires waitlist signup and approval[3] before usage. Also the PALM API has several caveats: first, it cannot translate non-English languages into English. Second, it sometimes refuses to return translations. Mostly the reasons for this filtering is unspecified, and sometimes because of safety filters against toxicity. The number of failures for each language pair is specified in Table 4. Interestingly the English-French pair has the most failures, and it is also worth noting that most of the toxicity-related filtering occurred for this pair, which seemingly resulted in this language suffering from the most filtering.

We prompted PALM2 in the following format: *Translate the following sentence from* $\{$*source language*$\}$ *to* $\{$*target language*$\}$*:* $\{$*text to translate*$\}$. Basically this is the same prompt as GPT-3.5 without the triple backticks, because we found that PALM2 would often echo the triple backticks in output.

As the PALM API only provides a subset of translations for the dataset, we use the same subset to measure performance when comparing it to other models. Table 3 provides a comparison of performance between PALM2, GPT-3.5, and Opus MT based on the subset of sentences the PALM2 is able to process. As can be seen in the table, both Opus MT and GPT-3.5 perform significantly worse on translations into Hindi, while PALM2 shows slightly worse performance compared to other language pairs. This is likely because PALM2 has been trained on a more multilingual, less English-centered dataset.

5.4 mBART-50

mBART-50 (Tang et al., 2020) is an extended version of mBART(Liu et al., 2020). It is finetuned on 50 languages and in *multiple* directions instead

[1] https://github.com/openai/openai-cookbook/blob/main/examples/api_request_parallel_processor.py

[2] https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/2/guidelines

[3] https://developers.generativeai.google/

Table 4: Number of failed translations per language pair for PALM2

language pair	Number of failures
English-German	13
English-French	24
English-Czech	9
English-Hindi	9
English-Russian	12

of only bilingual directions. mBART concatenates monolingual documents from each language and trains as a denoising autoencoder. We use the mBART-50 model available in the Transformers library to conduct our experiments. It took approximately four hours to evaluate each language pair on a Standard_D2S_v3 Azure instance.

Perhaps because it is a relatively older model, mBART-50 seems to perform the worst on most language pairs except for Hindi.

6 Error analysis & Performance

Since we were not fluent in any of the translated languages aside from English, we needed a method of evaluating the quality of translations. We looked into using BiLingual Evaluation Understudy (BLEU) scoring as our evaluation metric. BLEU score (Papineni et al., 2002) is a weighted geometric mean of n-gram precisions, multiplied by a brevity penalty. It is considered one of the best and inexpensive quality metrics, because it is highly correlated with evaluations performed by humans.

Unfortunately, BLEU score also has a number of shortcomings. It struggles with languages that have different or lacking word boundaries, and scores can wildly vary depending on the tokenization method. This makes it difficult to compare BLEU scores since varied parameters are commonly not reported. For this reason, we use the SacreBLEU (Post, 2018) variation. This is a version of BLEU that does not allow user-supplied parameters. All preprocessing is done internally.

In our case, we use only SacreBLEU on each of our translated outputs against the WMT14 reference datasets. It returns the score, which we have compiled in Table 1 and Table 2.

While we only use SacreBLEU in our case, a better indicator of translation quality would be to use something like COMET-22 (Rei et al., 2022), which is a reference-based metric that combines direct assessments (DA), sentence-level scores, and word-level tags in conjunction with BLEU. This should be done in future evaluations of machine translation models.

6.1 Comparisons of models

Generally, GPT-3.5 performed better than Opus MT on all language pairs except for English-Czech. Overall PALM2 seems to produce better translations on the subset data. Analyzing sentence pairs with the lowest BLEU scores showed that GPT-3.5 seems to be bad at translating certain sentences, regardless of the language. These are shown in table 5. They mostly seem to be newspaper article titles which are not full sentences.

One area in which GPT-3.5 prevails over OPUS-MT is in using the right pronouns and conveying the right ”meaning” rather than translating the literal words. Although both GPT-3.5 and OPUS-MT are ultimately based on Transformers, GPT-3.5 may have been trained on a larger and more diverse dataset. Additionally, GPT-3.5 may be better at capturing broader context, and focuses on coherent response as opposed to a precise translation. Interestingly this sometimes results in GPT-3.5 outputting even more ”fluent” responses than the dataset's references (e.g. GPT-3.5: *Old principle: show, don't tell.* vs. Reference: *The old basis of showing not telling.*).

6.2 GPT-3.5

Similar to what Zhu et al. (2023) discover, GPT-3.5 performs worse at translations into non-English languages. However this problem does not seem to be limited to GPT-3.5, as we can observe in Table 2. Translations to and from European languages(English, Czech, French, Russian) are generally accurate, with most errors being related to fluency (refer to the first entry in Table 6). Interestingly many translations contain unneeded quotation marks.

Out of the 5 languages, French is especially interesting because all models report the highest BLEU scores on both sides of this language pair. Another unexpected discovery is that for French, the accuracy of translations into French from English is higher than vice versa. This is an interesting phenomena as the opposite is observed for all other language pairs: models are better at translating other languages into English than the other way around. French is the second most prevalent language GPT-3 (GPT-3.5's direct predeces-

Table 5: Examples of sentences with low BLEU scores for GPT-3.5

Obama's Health Care Walk Back
Court blocks ruling on NYPD stop-and-frisk policy
China plea paper 'to be overhauled'

sor) was trained on, with approximately 1.78% of the characters used to train it being French. Though specific details about GPT-3.5's training data is not public, we can hypothesize that it would have used similar training data–thus it is likely that GPT-3.5 is proficient at French because it has been trained on many French documents.

Hindi was the language pair which most models performed the worst on. It had a significantly higher number of wrong translations where the meaning of the original sentence was not correctly represented in translations. Sometimes this was likely due to the term never being seen before, which lead to cases such as translating the "NRIG" into the "National Research Institute". Other times it mistranslated more common words such as "kidnap" and "puppy", which may be due to GPT-3.5 not having enough exposure to Hindi data. This is also likely because Hindi is different from English and French in that there are not explicit spaces between words. This means that although the the model may know of these words, it may not be able to pick up on them when tokenizing. It is also possible that zero-shot learning does not work so well for non-European languages.

6.3 PALM2

As the PALM API did not provide translations into English, it was difficult for us to perform manual error analysis on it. Instead, we focus on analyzing the quantitative performance of PALM2. As can be observed in Table 3 PALM2 greatly outperforms other models on Hindi, which is likely a result of its focus on non-English, low-resource languages. However it seems to lag behind on French and Czech translations compared to Opus MT–we believe this may be due to negative interference, where being trained on multiple languages results in worse performance on high-resource (European) languages. We used google translate APIs to attempt to perform manual error analysis after translating languages to English, but ran into issues with rate-limiting. Also, we would be unsure whether performance quality came from PALM, or from google translate.

After manual analysis of the sentences that the PALM API refused to translate, we discover that most of them contain the word "sex", such as *intersex*, *sex abuse*, and *sexual assault*. As mentioned by Anil et al. (2023), PALM2 was developed with much more precautions against toxicity in mind, so this is somewhat expected behavior.

6.4 Opus-MT

Opus-MT generally performs worse on every language compared to GPT-3.5. Although both models are Transformer-based, this is likely because GPT-3.5 is a stronger model with hundreds of billions of parameters. Meanwhile, Opus-MT only has tens to hundreds of millions of parameters. Additionally, GPT-3.5 was not only intended for machine translation; its layers of Transformers and wide variety of training enables it to provide fluent outputs even though it may not always give the best translations.

Opus-MT performs slightly better than mBART-50 for several reasons. The first is that it contains one model for each language pair. mBART-50, on the other hand, is a truly multilingual model which is trained on 50 languages itself. This limits the parameters in mBART-50, although it is a relatively old model. Newer multilingual models such as GPT-3.5 and M2M100 fare better.

6.4.1 French-English

When translating from French to English, there are certain areas in which Opus MT falls short.

French, unlike English, assigns gender-specific pronouns to objects as well as people. Thus, the pronouns *il, elle* (English: he, her), are used when referring to objects and people. We notice that Opus-MT occasionally outputs "it" when translating to English, when the intended output is "he". This may be a difficult property to encode without surrounding context, and may require stronger training data that covers a wide range of context variations.

French also has reflexive verbs, something not typically found in English. They typically translate to the passive voice in English. For exam-

Table 6: Error Analysis Samples

No.	Model	Language Pair	Translation	Reference
1	GPT-3.5	en-de	Or they don't choose any device at all and instead pay a flat fee based on the average miles driven by all residents of the state.	Or they can choose not to have a device at all, opting instead to pay a flat fee based on the average number of miles driven by all state residents.
2	GPT-3.5	en-hi	Threats have been given to lift the daughter if payment is not made.	If no money was paid he was threatened and told his daughter would be kidnapped.
3	GPT-3.5	en-cs	Are you thinking about doing this in quotation marks full-time?	Do you think you could do that "full time"?
4	OPUS-MT	en-fr	Many see a solution in the form of a small black box that attaches itself above your car's dashboard.	Many are beginning to see a solution in a little black box that fits neatly by the dashboard of your car.
5	OPUS-MT	en-fr	Equipped with a wingsuit (a combination of wings), it went up to 160 km/h above the famous Monserrate sanctuary.	Wearing a wingsuit, he flew past over the famous Monserrate Sanctuary at 160km/h.

ple, the verb *s'attacher* means "to attach oneself" or "to be (physically) attached", depending on the context. The model often chooses the wrong meaning, as we see in sample 4 of Table 5.

7 Contributions of group members

List what each member of the group contributed to this project here. For example:

- Surya Auroprem: ran and scored OPUS-MT translations, did error analysis and some writing
- Heejoo Shin: created scripts to evaluate PALM2, GPT-3.5, and mBART-50, did error analysis and some writeup
- Rafferty Chen: researched ways to run LLaMA and LLaMa-based models locally and experimented with them to find one that could do translation (got results for some language pairs, but did not have good enough hardware to run everything), some writing
- Shanqing Wang: attempted PALM2 evaluation, surveyed error analysis methodology, did error analysis and some writing

8 Conclusion

It was significantly harder than we expected to get various MMT models to work. This was largely due to lack of computing resources such as GPUs and various problems encountered when trying to implement open source projects. If we could continue working on our project, we would likely try applying the models we explored to other datasets, and also try evaluating translations with metrics other than BLEU. Another point worth exploring would be LLM prompting. We did not have much prior experience with best practices for prompting LLMs, and the rate limit made it difficult to actively experiment with different kinds of prompts. It would also be intersting to see how different prompts may lead to better/worse quality in translations.

We had expected that LLMs such as PALM and GPT-3.5 would perform well on multilingual translations, but it was surprising to see how well Opus MT, a relatively small model that could run on our laptops, performed. We were also surprised that mBART performed worse than Opus MT on most language pairs except for Hindi. Overall we feel that despite the surprisingly high performance of LLMs and MMT models, they are still not able to beat state-of-the-art bilingual models. Another factor to consider is that LLMs are huge and therefore take much time and resources to train(GPT-3 is trained on 170 billion parameters, PALM2 has 340 billion parameters). Therefore while LLMs may be good for a wide array of purposes, if we only want to focus on high-quality machine translation we will likely get better translations from bilingual machine translation models at a lower cost.

9 Artificial Intelligence (AI) tools Disclosure

We used tools such as ChatGPT and Bard to gain starter code for testing different MMT models. While they did provide some code, it was mostly non-functional and required many extra hours of manual work going through documentation and fixing various errors before we gained code that actually worked.

References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. (2023). Palm 2 technical report.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting bleu scores.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. T. (2022). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2022). Prompting palm for translation: Assessing strategies and performance.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. (2020). On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., Li, L., and Huang, S. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.