Domain Terminology Integration into Machine Translation: Leveraging Large Language Models

Yasmin Moslem[1,2,§], Gianfranco Romani[3], Mahdi Molaei[4], Rejwanul Haque[1,5], John D. Kelleher[1,6], and Andy Way[1,2]

[1]ADAPT Centre

[2]School of Computing, Dublin City University, Dublin, Ireland

[3]Thomson Reuters, Zug, Switzerland

[4]Department of Computer Engineering, University of Tabriz, Tabriz, Iran

[5]Department of Computing, South East Technological University, Carlow, Ireland

[6]Hamilton Institute, Maynooth University, Maynooth, Ireland

Abstract

This paper discusses the methods that we used for our submissions to the WMT 2023 Terminology Shared Task for German-to-English (DE-EN), English-to-Czech (EN-CS), and Chinese-to-English (ZH-EN) language pairs. The task aims to advance machine translation (MT) by challenging participants to develop systems that accurately translate technical terms, ultimately enhancing communication and understanding in specialised domains. To this end, we conduct experiments that utilise large language models (LLMs) for two purposes: generating synthetic bilingual terminology-based data, and post-editing translations generated by an MT model through incorporating pre-approved terms. Our system employs a four-step process: (i) using an LLM to generate bilingual synthetic data based on the provided terminology, (ii) fine-tuning a generic encoder-decoder MT model, with a mix of the terminology-based synthetic data generated in the first step and a randomly sampled portion of the original generic training data, (iii) generating translations with the fine-tuned MT model, and (iv) finally, leveraging an LLM for terminology-constrained automatic post-editing of the translations that do not include the required terms. The results demonstrate the effectiveness of our proposed approach in improving the integration of pre-approved terms into translations. The number of terms incorporated into the translations of the blind dataset increases from an average of 36.67% with the generic model to an average of 72.88% by the end of the process. In other words, successful utilisation of terms nearly doubles across the three language pairs.

1 Introduction

The primary goal of the WMT 2023 Terminology Shared Task is to evaluate the ability of MT systems to accurately translate technical terminology.

The task aims to assess the extent to which MT models can utilise additional information regarding the translation of terminology. The shared task requires the participants to provide three translations, one without terms and the others with two individual sets of terms.

There have been several advancements in the area of MT domain adaptation, where an MT model is expected to follow the style and terminology of a certain domain or client (Chu et al., 2017; Kobus et al., 2017). Moreover, some researchers give special focus to terminology while training and fine-tuning MT systems (Dinu et al., 2019; Hu et al., 2019b; Haque et al., 2020; Michon et al., 2020; Nayak et al., 2023). However, forcing an MT model to adhere to certain terminology at inference time is among the most challenging aspects of MT. Hence, several researchers have investigated approaches to terminology-constrained decoding at translation time (Hokamp and Liu, 2017; Hasler et al., 2018; Post and Vilar, 2018; Hu et al., 2019a; Exel et al., 2020). The goal is to ensure that the MT system can accommodate unseen terminology while retaining translation accuracy and fluency.

Recently, since the emergence of advanced LLMs such as GPT-3 (Brown et al., 2020), BLOOM (Le Scao et al., 2022), PaLM (Chowdhery et al., 2022), Falcon (Penedo et al., 2023), Llama 2 (Touvron et al., 2023), and Jais (Sengupta et al., 2023) to mention just a few, researchers have been exploring the capabilities of these models for a number of tasks including MT (Bawden and Yvon, 2023; Hendy et al., 2023; Jiao et al., 2023; Moslem et al., 2023; Vilar et al., 2023). Some work investigates whether it is possible to utilise LLMs for terminology-constrained MT using a pre-defined glossary (Moslem et al., 2023) or even a dictionary (Ghazvininejad et al., 2023). They found the approach is generally effective in increasing the number of terms used in the translation, even for low-resource languages.

We highlight our key contributions with the systems that we submitted for the WMT 2023 Terminology Shared Task as follows:

- **LLMs for domain-specific data augmentation:** In our previous work (Moslem et al., 2022), we employed LLMs, namely GPT-J (Wang and Komatsuzaki, 2021) and mGPT (Shliazhko et al., 2022), to generate domain-specific datasets based on the target sentences in a small authentic dataset, then generated the source sentences with back-translation (Sennrich et al., 2016; Poncelas et al., 2019), and finally fine-tuned an encoder-decoder MT model on this data. In this work, we take a couple of steps forward by instructing an LLM, namely ChatGPT (Brown et al., 2020; Ouyang et al., 2022), to generate terminology-based bilingual synthetic data. In other words, the LLM will generate both the source and target sides of translation pairs, making sure the pre-approved target terms provided by the organisers are used in the translations.

- **LLMs for terminology-constrained MT and MT post-editing:** In our previous work, we utilised an LLM for translation and provided it with a list of terms to support in-context learning, which improved adherence to the required terminology at inference time (Moslem et al., 2023). We also investigated whether we could use an LLM for post-editing MT generated by other systems. In this work, we prompt ChatGPT to insert missing terms into translations generated by an encoder-decoder MT system. In other words, if some of the translations generated by a fine-tuned MT model still do not include the terms provided by the organisers, we feed these translations into an LLM, namely ChatGPT, instructing it to incorporate these terms while using the same translation.

2 Method

In our submissions to the WMT 2023 Terminology Shared Task, we followed these steps:

(i) Generate bilingual synthetic data based on the pre-approved terms, using an LLM, namely ChatGPT.

(ii) Fine-tune a generic model, OPUS (Tiedemann and Thottingal, 2020), on a mix of the terminology-based synthetic data generated in (i) and a randomly sampled portion of the original generic training data.

(iii) Generate translations of the dev, test, and blind datasets provided by the organisers with the fine-tuned model from (ii).

(iv) Apply terminology-constrained automatic post-editing using ChatGPT to incorporate missing terms into translations that do not yet include the required terminology.

2.1 Synthetic Data Generation

We used ChatGPT "gpt-3.5-turbo"[2] to generate bilingual sentence pairs, using the terms provided by the organisers. So, given a target term, the model was asked to generate multiple translation pairs, including both the source (e.g. German) and the target (e.g. English). For parameters of ChatGPT's API, we used *top_p* 1 and *temperature* values 0 and 0.3 to generate diverse outputs.

> Example prompt: Terminology-based generation
>
> Please use the "Federal Ministry of Science" to generate just 20 numbered sentences in German-English in one Python dictionary format.

To filter the generated data, we first removed duplicate sentences from the whole dataset, based on both the source and target. Then, we applied language detection of both sides of the data using *fastText*[3] and *pycld2*[4] libraries to ensure that the generated sentences were in our desired languages. We excluded any sentences whose scores were below a certain threshold, namely 0.9 for *fastText* and 90 for *pycld2*.

The filtering step removed less than 1% of the generated data. However, due to computational resource and time limitations, we could not use all the generated data. Table 1 reports the number of generated, filtered, and used translation pairs.

Initially, we only had the development and test datasets, so we used them for the German-to-English language pair. Later, when the organisers released the blind dataset, we used the development, test and blind datasets for the Chinese-to-English and English-to-Czech language pairs.

---

[2]The model "gpt-3.5-turbo" is a relatively efficient and cost-effective option, so we wanted to understand the quality we can achieve with it.

[3]https://fasttext.cc/docs/en/language-identification.html

[4]https://github.com/aboSamoor/pycld2

Lang	Raw	Filtered	Used
DE-EN	124,215	104,318	68,265
EN-CS	187,471	103,797	64,218
ZH-EN	90,538	72,695	49,001

Table 1: Terminology-based bilingual data generated by ChatGPT for fine-tuning the OPUS model

To assess the quality of the bilingual data generated by ChatGPT, we computed cross-entropy scores (Moore and Lewis, 2010) of the synthetic translation pairs based on the strong encoder-decoder MT model, NLLB-200 3.3B (Costa-jussà et al., 2022). For scoring, we used CTranslate2[5] (Klein et al., 2020) *score_batch()* method with the parameters *batch_type* "tokens" and *max_batch_size* 2024. We scored each synthetic translation pair generated by ChatGPT, and then calculated the average score for the whole dataset. Computing dual cross-entropy scores according to two inverse translation models trained on clean data is an effective method to evaluate data quality (Junczys-Dowmunt, 2018). Hence, we computed the scores of both directions of each language pair according to the multilingual MT model NLLB-200 3.3B because both directions are generated by ChatGPT. To produce a baseline for translation quality, we generated the translations of the same datasets using NLLB-200 3.3B for each language direction with *beam_size* 4, and then scored these translations with the same model. As the scores are in the form of negative log probabilities, we converted them to their exponential equivalents for readability, which are reported in Table 2. It is normal that the model NLLB-200 generates higher scores for its own translations; however, we wanted to know to what extent such scores are comparable to those of ChatGPT's synthetic translation pairs. According to the scores, the German↔English language pair had the most comparable quality, followed by Czech↔English, and Chinese ↔English language pairs.

Among the approaches that can be employed for assessing the quality of synthetic bilingual data is semantic similarity between the two sides of each translation pair (e.g. with mUSE (Yang et al., 2020)). However, the scoring approach that we previously described and used achieves a similar goal while comparing the quality of the synthetic bilingual data to the translation quality of a strong MT baseline model, namely NLLB-200 3.3B.

Lang	ChatGPT	NLLB	Diff.
DE-EN	0.59	0.68	0.09
EN-DE	0.56	0.64	0.08
Avg.	0.58	0.66	0.08
CS-EN	0.58	0.70	0.12
EN-CS	0.49	0.58	0.09
Avg.	0.54	0.64	0.10
ZH-EN	0.39	0.56	0.17
EN-ZH	0.09	0.34	0.25
Avg.	0.24	0.45	0.21

Table 2: Scores of translation pairs generated by ChatGPT based on the NLLB-200 3.3B model

2.2 Fine-tuning

Using the term-based synthetic bilingual data generated in the previous step, we fine-tuned encoder-decoder Transformer-based MT models (Vaswani et al., 2017). In particular, we fine-tuned OPUS MT models, with Hugging Face Transformers.[6] We applied mixed fine-tuning (Chu et al., 2017); in other words, we fine-tuned the baseline model with a mix of the terminology-based synthetic data generated from the previous step (cf. Section 2.1) and a randomly sampled portion of the original generic data used to train the OPUS baseline model. The numbers of segments taken from the OPUS generic data are as follows: CS: 372,928, DE: 419,881, ZH: 462,780. We over-sampled the synthetic terminology-based data to make it the same size as the used portion of generic data. The fine-tuning parameters are as follows: *train* = 0.9, *val* = 0.1, *batch_size* = 32, *learning_rate* = 2e-5, *accumulate_gradient* = 4, *weight_decay* = 0.01, *num_train_epochs* = 1, *max_input_length* = 256, *max_target_length* = 256. Finally, we used the fine-tuned model to generate translations for the development, test, and blind sets.

At first glance, the fine-tuning step might look redundant if the LLM can achieve the same translation quality directly, either via zero-shot translation or few-shot in-context learning (Moslem et al., 2023). However, domain-specific or terminology-based knowledge distillation (Treviso et al., 2023) from a massive LLM to a compact task-oriented MT model can help boost efficiency at inference time while enhancing domain adaptation and terminology adherence. Obviously, when authentic in-domain data is available, it can be used for fine-tuning instead of synthetic data for domain adaptation of the MT model. In production workflows,
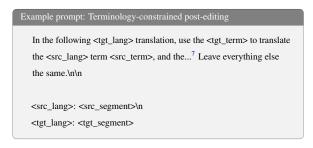
---

[5]https://github.com/OpenNMT/CTranslate2

[6]https://github.com/huggingface/transformers

only segments that do not meet specific quality criteria are passed to either human or automatic post-editing. Hence, deployment of a model fine-tuned on in-domain data can reduce the number of translations that need post-editing.

2.3 Terminology-constrained Automatic Post-Editing

For the shared task, the organisers provided two term sets for each source sentence in the test and blind datasets, and expected the participants to generate two translations that use one term set each. In this step of terminology-constrained automatic post-editing, we aim to refine the translations generated by an MT system by inserting the required terminology. To this end, we checked the translations generated by the fine-tuned model from the previous step (cf. Section 2.2). For each term set provided for the sentence, if the translation does not include all the terms, we ran this step of terminology insertion into the translation.

This step involves instructing ChatGPT to post-edit the translation by making sure it includes all the terms without changing the rest of the translation. For the API's parameters, we used *top_p* 1 and *temperature* values 0 and 0.2, and then chose the generation that fixed more terms.

Example prompt: Terminology-constrained post-editing

In the following <tgt_lang> translation, use the <tgt_term> to translate the <src_lang> term <src_term>, and the...[7] Leave everything else the same.\n\n

<src_lang>: <src_segment>\n

<tgt_lang>: <tgt_segment>

---

3 Evaluation

To assess the effectiveness of our process, we conducted two types of evaluation: (i) term-level evaluation in order to measure the level of adherence to the required terminology, and (ii) sentence-level evaluation in order to see whether the process affected the quality of the overall translation.

3.1 Term-level Evaluation

In Tables 3 and 4, we report the number of terms used in the translations of the test and blind datasets, respectively, in respect to the two term sets provided by the organisers. The results show the effectiveness of our proposed process, increasing the

---

[7]We can add more terms, if needed.

---

integration of the required terms in the final translations of the blind dataset from an average of 36.67% with the baseline generic model to an average of 72.88% after the LLM-based post-editing, across the three language pairs. Interestingly, prompting an LLM to integrate the required terms into the translations generated by a fine-tuned encoder-decoder MT model was more effective than solely using the fine-tuned model.

Lang	System	Total [1]	Used [1]	Total [2]	Used [2]	Avg %
DE-EN	Baseline	432	291	317	168	60.18
	Fine-tuned	432	302	317	165	60.98
	Term APE	432	**397**	317	**239**	**83.65**
EN-CS	Baseline	550	221	313	139	42.30
	Fine-tuned	550	135	313	108	29.53
	Term APE	550	**466**	313	**283**	**87.57**
ZH-EN	Baseline	1779	498	1938	491	26.66
	Fine-tuned	1779	854	1938	570	38.71
	Term APE	1779	**1137**	1938	**886**	**54.81**
Avg. %	Baseline					43.05
	Fine-tuned					43.07
	Term APE					**75.34**

Table 3: For the test dataset, the number of terms used in the translations from the first term set [1] and the second term set [2]. According to the results, terminology-constrained automatic post-editing ("Term APE") using ChatGPT achieved the best adoption of the required terminology.

Lang	System	Total [1]	Used [1]	Total [2]	Used [2]	Avg %
DE-EN	Baseline	11357	4120	11202	4623	38.77
	fine-tuned	11357	4130	11202	4621	38.81
	Term APE	11357	**6257**	11202	**5893**	**53.85**
EN-CS	Baseline	10626	3964	10563	5122	42.90
	Fine-tuned	10626	3397	10563	4412	36.87
	Term APE	10626	**8727**	10563	**8681**	**82.16**
ZH-EN	Baseline	2892	1375	2908	265	28.33
	Fine-tuned	2892	1422	2908	970	41.26
	Term APE	2892	**2471**	2908	**2322**	**82.65**
Avg. %	Baseline					36.67
	Fine-tuned					38.98
	Term APE					**72.88**

Table 4: For the blind dataset, the number of terms used in the translations from the first term set [1] and the second term set [2]. According to the results, terminology-based automatic post-editing ("Term APE") using ChatGPT achieved the best adoption of the required terminology.

3.2 Sentence-level Evaluation

After the end of the submission phase, the organisers released the references for the participants to conduct automatic evaluation. The main purpose of this sentence-based evaluation process is to determine whether terminology integration affected the overall quality of translation. In general, as demonstrated in Table 4 and Table 5, this terminology-constrained automatic post-editing step significantly increased the inclusion of the necessary

terms into the final translation while improving translation quality across the three language pairs.

For the automatic evaluation of each MT system, we used the BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET (Rei et al., 2020) metrics. Since many of the Chinese-to-English segments in the blind dataset did not have two term sets, we evaluated only those that had two term sets (1629 segments out of 2640 segments). We observe that the evaluation scores of the Chinese-to-English translation task are much lower than those of the two other language pairs. This can be due to the literary nature of the blind dataset extracted from Chinese novels, which might be difficult for both the MT model and automatic evaluation metrics.

Lang	Count	System	BLEU	chrF++	COMET
DE-EN	2963	Baseline	19.81	48.04	21.81
		Fine-tuned	19.27	47.75	21.51
		Term APE [1]	32.36	60.84	40.25
		Term APE [2]	27.84	56.84	33.20
		Term APE Avg.	**30.10**	**58.84**	**36.73**
EN-CS	3005	Baseline	29.13	53.11	50.90
		Fine-tuned	24.54	49.14	33.78
		Term APE [1]	45.65	67.36	79.84
		Term APE [2]	37.88	61.19	63.64
		Term APE Avg.	**41.77**	**64.28**	**71.74**
ZH-EN	1629	Baseline	6.95	27.95	-50.90
		Fine-tuned	7.76	29.26	-38.83
		Term APE [1]	9.56	32.80	-18.96
		Term APE [2]	11.93	35.30	-13.51
		Term APE Avg.	**10.75**	**34.05**	**-16.24**

Table 5: Automatic evaluation of the overall translation quality across the three language pairs based on the blind dataset. The "Baseline" refers to the OPUS model without fine-tuning, while "Fine-tuned" refers to the model after domain adaptation with the bilingual terminology-based synthetic data generated by an LLM. Finally, the three last rows for each language pair refer to using ChatGPT for terminology-constrained automatic post-editing ("Term APE") of the MT output generated by the fine-tuned model. In other words, "Term APE [1]" indicates the results when the first term set was used to prompt ChatGPT to integrate terms of this set into the translation generated by the fine-tuned model, while "Term APE [2]" refers to using the second term set. Finally, "Term APE Avg." is the average of "Term APE [1]" and "Term APE [2]" for each language pair. Terminology-constrained automatic post-editing with ChatGPT achieves the best results across the three language pairs in terms of the overall translation quality. As reported in Table 4, the number of terms integrated after the automatic post-editing step also increased.

Moreover, it is worth noting that we used the English term while generating bilingual synthetic data (cf. Section 2.1) for the three language pairs. However, English is the target language for both Chinese-to-English and German-to-English language directions, while it is the source language for the English-to-Czech language direction. This can explain the performance degradation after the

fine-tuning step in the English-to-Czech language direction (cf. Tables 4 and 5). In other words, it is recommended in the step of bilingual synthetic data generation to either use the target term or both the source and target terms while prompting the LLM to generate translation pairs.

As explained in Section 2.3, our final step of terminology-constrained automatic post-editing involves instructing an LLM to insert terms that were missing from the output of the fine-tuned model. This significantly increased term usage across all the Chinese-to-English, English-to-Czech, and German-to-English language pairs (cf. Table 4). Furthermore, as demonstrated in Table 5, this step had no detrimental effects on translation quality. In fact, integrating the necessary terms into the translation using ChatGPT improved translation quality according to our automatic evaluation.

4 Conclusion and Future Work

In this work, we showed that applying a multistep process of mixed fine-tuning on terminology-based synthetic bilingual data and then terminology-constrained automatic post-editing with an LLM can increase the adherence to the pre-approved terms in the generated translations. By the end of the process, the use of the required terms has increased in the translations of the blind dataset across the three language pairs from an average of 36.67% with the baseline generic model to an average of 72.88% after instructing an LLM to integrate the required terms into the translations.

Due to the task restrictions, we had to fine-tune OPUS models only. We would like to experiment with fine-tuning NLLB models, and probably the new SeamlessM4T (Barrault et al., 2023), Mistral (Jiang et al., 2023), and MADLAD-400 models (Kudugunta et al., 2023), on the same data and compare the output quality. In our experiments, we employed ChatGPT "gpt-3.5-turbo" for both terminology-based synthetic data generation and terminology-constrained automatic post-editing, as it is a relatively efficient and cost-effective option. In the future, we would like to repeat the same experiments with GPT-4 in order to assess the benefit of using a stronger language model on overall performance. We observe that BLOOM can be used as an alternative LLM for data generation; however, one-shot generation might work better than zero-shot generation. In this case, the prompt can consist of a term, a bilingual sentence pair,

and then another term. Interestingly, the model will predict a new translation pair including the second term. While certain open-source models such as Llama 2 and Falcon might be employed for the terminology-constrained automatic post-editing step for certain languages, we suspect that they will need fine-tuning before being reliably usable for most languages.

In future work, we will carry out a deeper analysis of the generated synthetic data together with the outputs of the fine-tuned models in order to understand how the properties of the synthetic data affect the fine-tuning results. It is important also to test the same approach for other languages, especially low-resource language pairs.

Moreover, it would be interesting to exclude the fine-tuning step and assess the overall translation quality after LLM-based post-editing. It is possible that domain adaptation through fine-tuning the baseline MT model either on authentic or synthetic data would still be beneficial. It can lead to domain-specific improvements in the overall translation quality that may not be achievable by the baseline model or the terminology-constrained post-editing step. Again, deploying a model fine-tuned on in-domain data into production can enhance terminology adherence in initial translations. As there is no need to send the translations that already include the pre-approved terms to the LLM for terminology-constrained post-editing, this can reduce the number of translations that require post-editing. Such an efficient workflow can allow us to save resources, and minimise latency at inference time. Similarly, there are potential advantages of employing an LLM for post-editing rather than for direct translation. Instead of solely relying on the translation quality of the LLM, quality estimation can be performed to select the best MT model in general or for the current source text segment. Ultimately, only segments that do not meet quality criteria are then passed to the LLM for post-editing.

Acknowledgements

References

Loic Barrault, Andy Chung, David Dale, Ning Dong (ai), Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng-Jen Chen, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Abinesh Ramakrishnan, Alexandre Mourachko, Amanda Kallet, Ann Lee, Anna Sun, Bapi Akula, Benjamin Peloquin, Bernie Huang, Bokai Yu, Brian Ellis, Can Balioglu, Carleigh Wood, Changhan Wang, Christophe Ropers, Cynthia Gao, Daniel Li (fair), Elahe Kalbassi, Ethan Ye, Gabriel Mejia Gonzalez, Hirofumi Inaguma, Holger Schwenk, Igor Tufanov, Ilia Kulikov, Janice Lam, Jeff Wang (pm Ai), Juan Pino, Justin Haaheim, Justine Kao, Prangthip Hasanti, Kevin Tran, Maha Elbayad, Marta R Costa-jussa, Mohamed Ramadan, Naji El Hachem, Onur Çelebi, Paco Guzmán, Paden Tomasello, Pengwei Li, Pierre Andrews, Ruslan Mavlyutov, Russ Howes, Safiyyah Saleem, Skyler Wang, Somya Jain, Sravya Popuri, Tuan Tran, Vish Vogeti, Xutai Ma, and Yilin Yang. 2023. SeamlessM4T—Massively Multilingual & Multimodal Machine Translation.

Rachel Bawden and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901, Virtual. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana

Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311 [cs.CL]*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672 [cs.CL]*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-Constrained Neural Machine Translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. *arXiv preprint arXiv:2302.07856 [cs.CL]*.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT's Submission to the Adap-MT 2020 English-to-Hindi AI Translation Shared Task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural Machine Translation Decoding with Terminology Constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210 [cs.CL]*.

Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019b. Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825 [cs.CL]*.

Wenxiang Jiao, Wenxuan Wang, Jen-Tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv preprint arXiv:2301.08745 [cs.CL]*.

Marcin Junczys-Dowmunt. 2018. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. Efficient and

high-quality neural machine translation with Open-NMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 372–378, Varna, Bulgaria.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *arXiv preprint arXiv:2309.04662 [cs.CL]*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Eliza-

beth Salesky, Sabrina J Mielke, Wilson Y Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán,

Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100 [cs.CL]*.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert C Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-Specific Text Generation for Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Prashanth Nayak, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Instance-Based Domain Adaptation for Improving Terminology Translation. In *Proceedings of Machine Translation Summit XIX: Research Track*, pages 222–231, Macau SAR, China. Association for Machine Translation in the Americas.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, pages 27730–27744, New Orleans, Louisiana, USA. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv preprint arXiv:2306.01116 [cs.CL]*.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of Machine Translation Models with Back-Translated Data Using Transductive Data Selection Methods. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing CICLing 2019: Computational Linguistics and Intelligent Text Processing*, pages 567–579, La Rochelle, France. Springer Nature Switzerland.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin,

and Eric Xing. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv preprint arXiv:2308.16149 [cs.CL]*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mGPT: Few-Shot Learners Go Multilingual. *arXiv preprint arXiv:2204.07580 [cs.CL]*.

Jörg Tiedemann and Santhosh Thottingal. 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288 [cs.CL]*.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H Martins, André F T Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient methods for natural language processing: A survey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. Github (mesh-transformer-jax).

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.