Exploring Multilingual Pretrained Machine Translation Models for Interactive Translation

Ángel Navarro[1] annamar8@prhlt.upv.es
Francisco Casacuberta [1,2] fcn@prhlt.upv.es
[1]PRHLT, Universitat Politècnica de València, Spain,
[2]ValgrAI - Valencian Graduate School and Research Network for Artificial Intelligence, Camí de Vera s/n, 46022 Valencia, Spain

Abstract

Pre-trained large language models (LLM) constitute very important tools in many artificial intelligence applications. In this work, we explore the use of these models in interactive machine translation environments. In particular, we have chosen mBART (multilingual Bidirectional and Auto-Regressive Transformer) as one of these LLMs. The system enables users to refine the translation output interactively by providing feedback. The system utilizes a two-step process, where the NMT (Neural Machine Translation) model generates a preliminary translation in the first step, and the user performs one correction in the second step–repeating the process until the sentence is correctly translated. We assessed the performance of both mBART and the fine-tuned version by comparing them to a state-of-the-art machine translation model on a benchmark dataset regarding user effort, WSR (Word Stroke Ratio), and MAR (Mouse Action Ratio). The experimental results indicate that all the models performed comparably, suggesting that mBART is a viable option for an interactive machine translation environment, as it eliminates the need to train a model from scratch for this particular task. The implications of this finding extend to the development of new machine translation models for interactive environments, as it indicates that novel pre-trained models exhibit state-of-the-art performance in this domain, highlighting the potential benefits of adapting these models to specific needs.

1 Introduction

Machine translation (MT) has become an integral part of modern communication, facilitating cross-border communication and interaction among people from different linguistic backgrounds. However, the effectiveness of MT depends heavily on the quality of the translation models and the techniques used for training them. Recently, pre-trained multilingual MT models such as mBART (multilingual Bidirectional Auto-Regressive Transformer) (Liu et al., 2020), mT5 (Xue et al., 2020), and XLM (Lample and Conneau, 2019) have emerged as powerful tools that can achieve state-of-the-art performance on various benchmark datasets. Now, we can obtain high-quality translations for a specific task or domain by fine-tuning them with a not large training data, which is very effective in using them in low-resource settings.

However, even with the high performance of these pre-trained models, there are still challenges in achieving accurate and fluent translations for all languages and domains (Toral, 2020). Interactive machine translation (IMT), which combines human intelligence with MT, has been proposed as a potential solution to address these challenges and ensure consistently

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 132–142
September 4–8, 2023, Macau SAR, China.

high-quality translations (Peris et al., 2017). IMT systems allow users to actively participate in the translation process by providing feedback to the machine, which generates a new translation that corrects the previous error. This process repeats until the machine generates a perfect translation, and the user validates it.

In this paper, we explore the use of mBART in an IMT environment. We aim to investigate and compare the effectiveness of using a pre-trained model like mBART, which assesses state-of-the-art results in a large set of translation tasks, with models we train from scratch for a specific domain using the OpenNMT-py toolkit (Klein et al., 2017). To achieve this, we design and implement an IMT system with a prefix-based protocol (Foster et al., 1997) that integrates mBART. In this protocol, firstly described by Foster et al. (1997), further developed by Alabau et al. (2013); Barrachina et al. (2009); Langlais et al. (2000), the user only corrects at each iteration the first error that he finds from left to right. This process repeats until the user validates the machine-generated translation. We also fine-tune mBART to compare our results with the two versions of it, with and without the fine-tuning. To compare the different models, we use different evaluation metrics that help us evaluate the effort the user has to perform during the translation session. Our results showed that although the mBART models obtained higher quality translation than ours, the user effort results are similar. We need to fine-tune the mBART model on the specific domain to achieve a better effort reduction. In order to achieve optimal translation quality, our models should produce the most accurate translations possible on the initial attempt. However, when operating in an IMT environment, the model must adapt to user feedback. Based on our findings, it can be concluded that while pre-trained models generally achieve better translation quality, training a MT model from scratch for a specific domain can lead to better generalization, which produces significant benefits in this field.

Our work presents several contributions to the field of IMT. Specifically, our contributions are as follows:

- **Creation of an IMT system with mBART**: We have implemented an IMT system that uses as the principal MT model mBART. It uses a prefix-based protocol and forces the decoder to use the prefix that the user has validated.

- **Fine-Tune mBART**: The primary objective of our study is to investigate whether pre-trained models, known for achieving state-of-the-art results in translation tasks, can also perform well in the context of IMT. To conduct a more comprehensive experiment, we fine-tuned the model on a specific domain.

- **Compare IMT results with traditional techniques**: We compare and analyze the results obtained with the mBART models with ours, which have been trained from scratch. Our study evaluates the quality of translations and the level of user effort required during translation sessions.

The rest of the paper is organized as follows. In Section 2, we provide a brief overview of related work in pre-trained language models and IMT. Section 3 describes our proposed approach in detail, including the architecture of the IMT system and the feedback mechanism. In Section 4, we present the experimental framework, and in Section 5, we discuss the results obtained. Finally, we conclude the paper in Section 6 and discuss potential future directions for this research.

2 Related Work

In this work, our primary focus is to investigate whether pre-trained models, which have shown significant success in various tasks such as translation, can also be used effectively for IMT. Training a MT model from scratch for a specific domain can be time-consuming and challenging to obtain a suitable dataset. Therefore, evaluating whether pre-trained models can achieve similar results in this field as in MT tasks would be beneficial. Although MT and IMT tasks are similar, the goal of the first is to obtain the most accurate translation, while the second tries to obtain a perfect translation with minimal user interaction, so we need a model that can generalize well more than one that performs better translations.

This paper uses the multilingual pre-trained model mBART to compare its performance with our models trained from scratch for the specific task. There are other pre-trained multilingual models, such as mT5 (Xue et al., 2020), XLM, (Lample and Conneau, 2019) DeltaLM (Ma et al., 2021), or XGLM (Lin et al., 2021), that we could have used for our purpose. As we have used, other people are trying to use these models for new tasks that were not initially thought. Shen et al. (2021) used to resolve math word problems, Farahani et al. (2021) to summarize Persian texts, Chakrabarty et al. (2021) generated poetry with them, and Li et al. (2020) implemented it in the speech translation field.

Apart from pre-trained multilingual models, the advent of Large Language Models has prompted research into their utilization for specific tasks, including translation. These models have undergone extensive training on large-scale multilingual datasets, enabling them to capture linguistic patterns and translations across multiple languages (Scao et al., 2022; Hoffmann et al., 2022; Brown et al., 2020). In some scenarios, the translations produced by these models exhibit such remarkable quality that they pose a competitive challenge to the existing state-of-the-art translation models (Hendy et al., 2023; Zhang et al., 2023).

The task in which we have employed and compared mBART is that of IMT. This field has been under investigation since Foster et al. (1997), with the first appearance of the prefix approach, and has continued to evolve ever since. Numerous research branches have emerged, exploring different techniques to reduce human effort in translation. Domingo et al. (2017) proposed a fresh approach to the behavior of translators, transitioning from correcting at the prefix level to enabling the validation of multiple segments within a single translation. Other techniques aim to minimize human effort more directly, such as optimizing the utilization of user mouse actions (Navarro and Casacuberta, 2021b; Sanchis-Trilles et al., 2008) or implementing a confidence measurement system to provide an initial evaluation of the translation (Navarro and Casacuberta, 2021a; González-Rubio et al., 2010). Additional techniques take advantage of the ability of the IMT system to guarantee perfect translations, utilizing them to enhance the translation model through active and online learning techniques (Peris and Casacuberta, 2019, 2018; Rubio and Casacuberta, 2014). Some frameworks like Casmacat (Alabau et al., 2013) and TransType (Cubel et al., 2003) add a large set of these innovations in the same workplace. Commercial environments like *Lilt* and *Unbabel* can also use interactive machine translation.

In the upcoming section, we will explore the framework of prefix-based IMT, which will provide a deeper understanding of how we have tailored the mBART model for this specific task.

3 IMT Framework

First, it is essential to examine the neural machine translation (NMT) framework to elucidate the modifications undertaken to adapt it for IMT systems. The NMT framework, introduced by Castaño and Casacuberta (1997), has demonstrated its efficacy and power in recent years. Its impact and effectiveness have been widely recognized in the field of MT (Stahlberg, 2020; Klein et al., 2017). Given the sentence $x_1^J = x_1, ..., x_J$ from the source language X, to get the translation with the highest probability $\hat{y}_1^{\hat{I}} = \hat{y}_1, ..., \hat{y}_{\hat{I}}$ from the target language Y, the fundamental equation of the statistical approach to NMT would be:

$$\hat{y}_1^{\hat{I}} = \underset{I, y_1^I}{\arg\max}\ \Pr(y_1^I \mid x_1^J) = \underset{I, y_1^I}{\arg\max} \prod_{i=1}^{I} \Pr(y_i \mid y_1^{i-1}, x_1^J) \tag{1}$$

where $\Pr(y_i \mid y_1^{i-1}, x_1^J)$ is the probability distribution of the next word given the source sentence and the previous words. The distinguishing feature of the IMT framework lies in its utilization of human feedback as valuable information for determining the translation with the highest probability. In this framework, the professional users provide feedback when encountering the first error reviewing from left to right. When an error is identified at position p, the user moves the cursor to that position, producing the feedback $f_1^p = f_1, ..., f_p$, where f_1^{p-1} is the validated prefix, and f_p is the word that the user has typed to correct the error. The following equation adds this feedback to Equation 1 with two constraints that apply for the range of words $1 \leq i < p$:

$$\begin{array}{ll} \hat{y}_1^{\hat{I}} = \underset{I, y_1^I}{\arg\max}\ \Pr(y_1^I \mid x_1^J, \bar{y}_1^{\bar{I}}, f_1^p) = \underset{I, y_1^I}{\arg\max} \prod_{i=1}^{I} \Pr(y_i \mid y_1^{i-1}, x_1^J, \bar{y}_1^{\bar{I}}, f_1^p) & \\ \text{subject to} \quad 1 \leq i < p & \\ \qquad\qquad\quad f_i = y_i = \bar{y}_i & \\ \qquad\qquad\quad f_p = y_p \neq \bar{y}_p & \end{array} \tag{2}$$

where $\bar{y}_1^{\bar{I}} = \bar{y}_1, ..., \bar{y}_{\bar{I}}$ is the previous translation, f_1^p is the feedback provided by the user, which corresponds with the validated prefix with the new word typed, and p is the length of the feedback. With constrain $f_i = y_i = \bar{y}_i$, we assure that all the words before the error position, the validated prefix, are in the new translation, and with constrain $f_p = y_p \neq \bar{y}_p$, we force to use the new word typed by the user. As the user corrects and validates the translation from left to right, in a more general way, this equation generates the most probable suffix for the prefix provided.

We have implemented this framework to be compatible with models from both *OpenNMT-py* toolkit (Klein et al., 2017), which we trained from scratch, and HuggingFace library (Wolf et al., 2020), from where we obtained the mBART model checkpoint. We have developed an IMT system in which the translation model interacts with a simulated user to generate translations. The simulated user detects the first error by comparing the generated translation with the reference word-by-word. Section 4.4 of the paper describes the simulation process in more detail.

4 Experimental Framework

This section provides a comprehensive account of our experimental procedures, beginning with an overview of the evaluation metrics used to assess our proposal. We then describe the corpora utilized to train and test our models and outline the specific training procedures employed for our machine translation systems. Finally, we describe the user simulation process in detail.

4.1 Evaluation metrics

We made use of the following well-known metrics in order to assess our proposal:

Word stroke ratio (WSR) Tomás and Casacuberta (2006): measures the number of words typed by the user, normalized by the number of words in the final translation.

Mouse action ratio (MAR) Barrachina et al. (2009): measures the number of mouse actions made by the user, normalized by the number of characters in the final translation.

		Europarl		
		De–En	**Es–En**	**Fr–En**
Train	$\|S\|$	1.9M	2.0M	2.0M
	$\|T\|$	49.8M/52.3M	51.6M/49.2M	60.5M/54.5M
	$\|V\|$	394.6K/129.1K	422.6K/309.0K	160.0K/131.2K
Val.	$\|S\|$	3000	3003	3000
	$\|T\|$	63.5K/64.8K	69.5K/63.8K	73.7K/64.8K
	$\|V\|$	12.7K/9.7K	16.5K/14.3K	11.5K/9.7K
Test	$\|S\|$	2169	3000	1500
	$\|T\|$	44.1K/46.8K	62.0K/56.1K	29.9K/27.2K
	$\|V\|$	10.0K/8.1K	15.2K/13.3K	6.3K/5.6K

Table 1: Corpora statistics. K denotes thousands and M millions. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. **Fr** denotes French; **En**, English; **De**, German; and **Es**, Spanish.

Additionally, we assessed the initial translation quality of each system using:

Bilingual evaluation understudy (BLEU) Papineni et al. (2002): computes the geometric average of the modified n-gram precision, multiplied by a brevity factor that penalizes short sentences. In order to ensure consistent BLEU scores, we used *sacreBLEU* Post (2018) for computing this metric.

Translation error rate (TER) Snover et al. (2006): computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation. It can be seen as a simplification of the user effort of correcting a translation hypothesis on a classical post-editing scenario.

4.2 Corpora

For our experiments, we utilized the Europarl corpus, which is a collection of proceedings from the Europarl Parliament. We used the training set to train our MT model and to fine-tune mBART. To validate and test the De–En and Fr–En models, we used WMT[1][2]'s's *news-test2013* and *news-test2015* datasets, respectively. For the Es–En models, we used *news-test2012* and *news-test2013* for validation and test purposes. It is worth noting that these datasets are commonly used in machine translation research and provide a benchmark for evaluating the performance of the models.

Table 1 shows the main features of the corpora.

4.3 Systems

Our system was built using the *OpenNMT-py* toolkit (Klein et al., 2017) and employed a Transformer architecture (Vaswani et al., 2017) that consisted of 6 layers, with all dimensions set to 512 except for the hidden Transformer feed-forward layer, which was set to 2048. We utilized 8 heads of Transformer self-attention, with 2 batches of words in a sequence to run the generator in parallel. A dropout of 0.1 was applied, and the optimization was carried out using Adam (Kingma and Ba, 2017) with a beta2 of 0.998, a learning rate of 2, and Noam learning rate decay with 8000 warm-up steps. We implemented label smoothing of 0.1 (Szegedy et al.,

[1] `http://www.statmt.org/wmt12/translation-task.html`.

[2] `http://www.statmt.org/wmt15/translation-task.html`.

SOURCE: El Estado de Indiana fue el primero en exigirlo.
TARGET: Indiana was the first State to impose such a requirement.

ITER-0	Translation hypothesis	Indiana was the sooner State to impose that condition.
ITER-1	Feedback	*Indiana was the* **first**
	Translation hypothesis	*Indiana was the first* State to impose such a condition.
ITER-2	Feedback	*State to impose such a* **requirement**
	Translation hypothesis	*Indiana was the first State to impose such a requirement*.
END	Final translation	*Indiana was the first State to impose such a requirement.*

Figure 1: Prefix-based IMT session to translate a sentence from Spanish to English. The process starts with the system offering an initial hypothesis. Then, at iteration 1, the user makes a word correction (**first**), validating the prefix *Indiana was the*. The system reacts to this feedback by generating a new translation hypothesis. Once more, the user reviews the hypothesis, making the word correction **requirement**, and updating the validated prefix *Indiana was the first State to impose such a*. Finally, since the next hypothesis is the desired translation, the process ends with the user accepting the translation. Overall, this process has a post-editing effort of 2 wordstrokes and 3 mouse actions.

2015) and used beam search with a beam size of 6. Finally, we applied joint byte pair encoding to all corpora, merging them using $32,000$ operations.

We utilized the *facebook/mbart-large-50-many-to-many-mmt*[3] checkpoint (Tang et al., 2020) from the Hugging Face library (Wolf et al., 2020), which employs a Seq2Seq Transformer architecture (Vaswani et al., 2017). The model consists of 12 encoder layers and 12 decoder layers, with a model dimension of 1024 and 16 heads. For customizing the mBART model to a specific pair of languages of our domain, we fine-tuned it on a single bi-text dataset from the training set of the corpora, calling this models mBART FT. During training, we inputted the source language into the encoder and used the decoder to decode the target language, resulting in a new model for each language pair. We conducted 100K training updates with a learning rate of $2e-5$ and a weight decay of 0.01 for training each model.

4.4 Simulation

To minimize the high time and economic costs associated with frequent human evaluations during the development stage, we opted to use simulated users to conduct the evaluations. Additionally, due to the novelty of our work, as it represents the first integration of mBART with an IMT system, we decided to perform the experiments in a more controlled environment with the simulation. These simulated users were tasked with generating translations from a given reference.

To conduct these evaluations, we utilized the prefix-based protocol described by Foster et al. (1997), in which the user identifies and corrects the leftmost incorrect word, validating all the previous words in the prefix up to the point of correction. In other words, the validated prefix consists of all the words before and including the corrected word.

We have opted for this prefix-based protocol to assess the effectiveness of the LLMs in an IMT system by its simplicity. The post-edition work is more in line with the segment-based protocol in which the corrections are made throughout the translation, focusing on the most incorrect words first. This protocol may introduce additional variables that may alter the intended demonstration of this paper, gauge the ability to recover from errors and generate a correct translation of the LLM, thereby examining their effectiveness in the context of interactive

[3] `https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt`

translation. For this reason, as it helps us to isolate and evaluate the generative capacity of the LLMs, we have opted for the prefix-based protocol.

When the simulation begins, the system generates an initial hypothesis for the translation, which the simulated user then reviews. The user searches for the first error in the translation by comparing the words and their positions in the hypothesis with those in the reference. If the user identifies an error, they consult the reference to determine the correct word and provide it as feedback to the system. This feedback is entered into the system by performing a word stroke, and if the error position is not adjacent to the previous correction, a mouse action is also required. This process is repeated until the simulated user has successfully translated the entire sentence without any errors. The user performs a mouse action to validate the translation, indicating that the entire sentence has been correctly translated.

Figure 1 presents an example of the simulation performed to translate a source sentence. The translation session starts with the system generating an initial hypothesis that needs to be reviewed and corrected. Then, at iteration 1, the simulated user corrects the word *first* at the fourth position, validating all the previous words. With the feedback provided, the system generates a new hypothesis. At iteration 2, the simulated user corrects the word *requirement*, updating the validated prefix. This time the translation hypothesis that the system generates with the new feedback is correct, and the simulated user validates at the next iteration.

5 Results

In order to perform our comparison, we evaluated the three different models in a prefix-based IMT system. We aim to see a reduction in human effort when using the mBART models. In this case, we evaluate the human effort with two different metrics, the WSR, and the MAR, each evaluating a different kind of effort—the effort performed by typing the words and the one by moving the mouse. In order to achieve a more precise evaluation of the models, we will consider that the effort required to type a word exceeds that of moving the mouse. Furthermore, in a professional setting with real translators, there may be instances where mouse actions are performed using the keyboard, diminishing their implication.

Table 2 shows the experimental results, where the *OpenNMT-py* model trained from scratch is compared with mBART and mBART FT. The quality of the models in terms of TER and BLEU is included for each experiment to get a grasp of the quality of the initial hypothesis that the simulated users will translate interactively with the IMT system. The first observation that can be made when looking at the results is that the best MAR values were obtained in all experiments using the *OpenNMT-py* model. However, it is important to note that this happens while not consistently achieving the lowest WSR values due to the fact that the errors identified by the simulated user are contiguous, thereby eliminating the need for mouse movement to correct each subsequent error. This hypothesis is supported by the TER values obtained, which despite not producing translations of the highest quality in terms of BLEU, indicate that the translations generated need a lower number of word editing operations, suggesting that errors within a translation are grouped. This fact also means that the translations generated with the mBART models, while producing higher-quality translations, have their errors more distributed.

Regarding the effort derived from keyboard usage, evaluated through the WSR and assumed to be more important than mouse usage, the mBART models achieve the best results. Except for the En–De language pair experiment, the mBART FT model has successfully reduced the effort compared to our baseline model, suggesting that proper fine-tuning of mBART can yield better results in the field of interactive machine translation than training a model from scratch. It is also worth noting that when comparing the *OpenNMT-py* model with mBART, in cases where the target language was English, mBART achieved superior results without requiring fine-tuning. This fact aligns with the findings reported in Tang et al. (2020), where the best results were

		Translation Quality		**User Effort**	
Model	**Language Pair**	**TER [↓]**	**BLEU [↑]**	**WSR [↓]**	**MAR [↓]**
OpenNMT-py	De–En	60.91	20.67	38.5	**4.6**
	En–De	66.31	17.35	**34.9**	**4.2**
	Es–En	70.56	15.90	48.3	**4.8**
	En–Es	**57.05**	23.88	37.4	**4.3**
	Fr–En	**54.94**	25.83	37.3	**4.6**
	En–Fr	**55.33**	32.16	35.5	**4.1**
mBART	De–En	**58.55**	**31.69**	**35.2**	7.0
	En–De	65.50	27.56	37.6	6.5
	Es–En	65.82	**31.09**	38.5	6.8
	En–Es	64.27	**29.66**	39.3	6.8
	Fr–En	57.56	**34.17**	34.9	7.4
	En–Fr	62.40	24.90	40.2	8.1
mBART FT	De–En	60.49	30.49	36.7	5.8
	En–De	**64.94**	**27.92**	36.8	5.2
	Es–En	**61.14**	31.03	**36.3**	5.8
	En–Es	58.86	33.47	**35.3**	5.3
	Fr–En	57.67	34.00	**34.7**	5.8
	En–Fr	57.87	**40.07**	**30.7**	5.4

Table 2: Results of the *OpenNMT-py*, mBART, and mBART FT models in a prefix-based IMT system. All values are reported as percentages. Best results are denoted in bold.

achieved in the Many-to-One configuration when translating into English.

In summary, the best results in terms of WSR for reducing the human effort during interactive machine translation sessions in a prefix-based environment have been achieved using the mBART FT model, which has shown reductions in WSR of up to 5 points. This indicates that if maximum effort reduction is desired, fine-tuning the model to our specific domain is necessary. For tasks targeting the English language, the base mBART model has already demonstrated a reduction in human effort, suggesting that for such tasks, using the base mBART model may be more beneficial and efficient than training a model from scratch.

6 Conclusions and future work

In this study, we have compared the effectiveness of pretrained multilingual machine translation models with those we can train from scratch in the IMT field. Both models have achieved similar results, although mBART has excelled in language pairs where the target language is English. Furthermore, by fine-tuning the pretrained models in the specific domain, the reduction in human effort is further improved, surpassing our baseline model. This confirms that pretrained models can also yield good results in this field after adjusting the model for the specific domain. By performing fine-tuning instead of training a translation model from scratch, we can significantly reduce the computational cost associated with training. This approach allows us to achieve a competent model while minimizing computational resources.

Based on the obtained results, we can conclude that mBART with fine-tuning achieves better results in the field of IMT compared to training a model from scratch. As future work, it would be interesting to investigate whether other pretrained models, such as mT5, exhibit similar characteristics. Additionally, conducting a comparative analysis among these pretrained models would provide valuable insights.

Acknowledgements

This work received funding from *Generalitat Valencia* under the program *CIACIF/2021/292* and from *ValgrAI* (*Valencian Graduate School and Research Network for Artificial Intelligence*). Work partially supported by grant PID2021-124719OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by *European Regional Development Fund* (*ERDF*).

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González-Rubio, J., Koehn, P., Leiva, L., Mesa-Lao, B., et al. (2013). Casmacat: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100(1):101–112.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Castaño, A. and Casacuberta, F. (1997). A connectionist approach to machine translation. In *Fifth European Conference on Speech Communication and Technology*, pages 91–94.

Chakrabarty, T., Saakyan, A., and Muresan, S. (2021). Don't go far off: An empirical study on neural poetry translation. *arXiv preprint arXiv:2109.02972*.

Cubel, E., González, J., Lagarda, A., Casacuberta, F., Juan, A., and Vidal, E. (2003). Adapting finite-state translation to the transtype2 project. In *EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT*, pages 15–17, Budapest, Hungary. European Association for Machine Translation.

Domingo, M., Peris, Á., and Casacuberta, F. (2017). Segment-based interactive-predictive machine translation. *Machine Translation*, 31(4):163–185.

Farahani, M., Gharachorloo, M., and Manthouri, M. (2021). Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–6. IEEE.

Foster, G., Isabelle, P., and Plamondon, P. (1997). Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194.

González-Rubio, J., Ortíz-Martínez, D., and Casacuberta, F. (2010). Balancing user effort and translation error in interactive machine translation via confidence measures. In *Proceedings of the Association for Computational Linguistics 2010 Conference Short Papers*, pages 173–177, Uppsala, Sweden. ACL.

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. *arXiv preprint arXiv:2302.09210*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the Association for Computational Linguistics 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291.*

Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: a computer-aided translation typing system. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*, pages 46–51.

Li, X., Wang, C., Tang, Y., Tran, C., Tang, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2020). Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829.*

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2021). Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

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.

Ma, S., Dong, L., Huang, S., Zhang, D., Muzio, A., Singhal, S., Awadalla, H. H., Song, X., and Wei, F. (2021). Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Navarro, Á. and Casacuberta, F. (2021a). Confidence Measures for Interactive Neural Machine Translation. In *Proceedings of the IberSPEECH 2021*, pages 195–199. IberSPEECH.

Navarro, Á. and Casacuberta, F. (2021b). Introducing mouse actions into interactive-predictive neural machine translation. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 270–281.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. ACL.

Peris, Á. and Casacuberta, F. (2018). Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. ACL.

Peris, Á. and Casacuberta, F. (2019). Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language*, 58:98–126.

Peris, Á., Domingo, M., and Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191.

Rubio, J. G. and Casacuberta, F. (2014). Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters*, 37:124–134. Partially Supervised Learning for Pattern Recognition.

Sanchis-Trilles, G., Ortíz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., and Hoang, H. (2008). Improving interactive machine translation via mouse actions. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 485–494, Honolulu, Hawaii. Association for Computational Linguistics.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shen, J., Yin, Y., Li, L., Shang, L., Jiang, X., Zhang, M., and Liu, Q. (2021). Generate & rank: A multi-task framework for math word problems. *arXiv preprint arXiv:2109.03034*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. AMTA.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

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. *arXiv preprint arXiv:2008.00401*.

Tomás, J. and Casacuberta, F. (2006). Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 835–841, Sydney, Australia. ACL.

Toral, A. (2020). Reassessing claims of human parity and super-human performance in machine translation at wmt 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. EAMT.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.