Setting up the Data Printer with Improved English to Ukrainian Machine Translation

Yurii Paniv, Dmytro Chaplynskyi, Nikita Trynus, Volodymyr Kyrylov

Ukrainian Catholic University, lang-uk initiative,
Igor Sikorsky Kyiv Polytechnic Institute, Università della Svizzera italiana
paniv@ucu.edu.ua, chaplinsky.dmitry@gmail.com, trynus.nikita@lll.kpi.ua, vol@wilab.org.ua

Abstract

To build large language models for Ukrainian we need to expand our corpora with large amounts of new algorithmic tasks expressed in natural language. Examples of task performance expressed in English are abundant, so with a high-quality translation system our community will be enabled to curate datasets faster. To aid this goal, we introduce a recipe to build a translation system using supervised finetuning of a large pretrained language model with a noisy parallel dataset of 3M pairs of Ukrainian and English sentences followed by a second phase of training using 17K examples selected by k-fold perplexity filtering on another dataset of higher quality. Our decoder-only model named Dragoman beats performance of previous state of the art encoder-decoder models on the FLORES devtest set.

Keywords: machine translation, parameter-efficient fine tuning, large language models, unsupervised data selection, perplexity filtering

1. Introduction

The availability of the data is the most important ingredient when one needs to pretrain general-purpose large language models for a specific natural language task or a set of tasks. While it is relatively easy to obtain a good and balanced dataset under specific domain for the English language, it is much harder to do the same for other under-resourced languages such as Ukrainian.

Since curating a corpus of tasks in Ukrainian is a large endeavor, and given a large body of work done for English, we consider existing instruction tuning datasets as a source of tasks to reuse in Ukrainian using automatic machine translation.

This work focuses on improving the current state of machine translation from English to Ukrainian.

We contribute a recipe for finetuning a large pretrained language model with publicly available data to build a translation system (section 3, section 4). This matches state of the art performance of the best encoder-decoder model on a common multilingual benchmark using a consumer GPU with 24 GiB of VRAM. We release training, evaluation code, datasets, and model at `https://github.com/lang-uk/dragoman`. Our main results are summarized in Table 1. We provide examples of the top-5 best and worst translations on the FLORES devtest set in the Appendix A.

We base pretrained model selection on evaluation in few-shot learning setting (section 5). We find that it's a promising method to design tasks without training, and the model can perform comparably to specialized systems given increased inference budget and auxiliary translation scoring functions, yet still underperforms our finetuned recipe.

Model	BLEU ↑
Finetuned	
Dragoman P, 10 beams (section 3)	30.4
Dragoman PT, 10 beams (section 4)	**32.3**
Zero shot and few shot (section 5)	
Llama 2 7B 2-shot, 10 beams	20.1
Mistral-7B-v0.1 2-shot, 10 beams	24.9
gpt-4 10-shot	29.5
gpt-4-turbo-preview 0-shot	30.4
Pretrained encoder-decoder	
NLLB-3B, 10 beams	30.6
OPUS-MT, 10 beams	**32.2**

Table 1: Main results. Our Dragoman models improve existing state of the art on translation from English to Ukrainian on FLORES-101 devtest (Goyal et al., 2022), a multilingual benchmark of translated sentences from web articles. We compare to state of the art encoder-decoder models, NLLB-3B (Team et al., 2022) and OPUS-MT (Tiedemann and Thottingal, 2020).

2. Supervised Finetuning

We cast machine translation as a likelihood maximization of a density p of Ukrainian sentences $Y =$ "перекладене речення" $\in \mathcal{Y}$ conditioned on their English sources with quasi-instruction formatting: $X =$ "[INST] translated sentence [/INST]" $\in \mathcal{X}$.

The density is parametrized using a neural network with frozen pretrained weights θ:

$$\mathrm{argmax}_\phi \, \mathrm{p}_{\theta,\phi}(Y|X) \qquad (1)$$

We implement the conditional language modeling objective by masking out tokens of X when

Dataset	Pairs	Filters				Example Order	Best BLEU ↑
		Lang	BPC	LaBSE	Len diff		
1m unfiltered	963k	-	-	-	-	Random	28.26
1m filtered	958k	En/Uk	<3.33	>0.91	<50	Random	29.47
3m filtered	2.9m	En/Uk	<3.25	>0.85	<50	By LaBSE score, dissimilar first	**30.37**
8m filtered	8m	En/Uk	<5	>0.5	<50	By LaBSE score, dissimilar first	30.19

Table 2: Summary of experiments with Paracrawl subcorpora. Legend of filters: **Lang** denotes language filters, **BPC** denotes maximum sum of bits per character measures, **LaBSE** denotes maximum sentence embedding cosine similarity between source and target sentences, **Len diff** denotes maximum difference in length between source and target in characters. Example ordering impacts data loading in the training loop.

computing token-wise cross entropy of shifted targets. We only optimize extra low rank adapter (Hu et al., 2022) parameters ϕ after nf4 quantization (Dettmers et al., 2023). In practice we use large rank values and adapter mixture weights. All training runs proceed for one epoch and we use dropout (Srivastava et al., 2014) for regularization against data noise.

We use Mistral-7B-v0.1 (Jiang et al., 2023) as a base pretrained decoder-only transformer, as it performs favorably in our few-shot experiments (section 5).

3. First Phase: Heuristic Filtering of Paracrawl

We use the publicly available Paracrawl dataset (Bañón et al., 2020). This dataset contains 13,354,365 English-Ukrainian sentence pairs, collected by automatically matching similar sentences in large corpora of internet text.

We have identified issues with translation pairs, including a significant number of repetitive or incorrect examples. We encounter a large subset of repetitive weather forecasts following the template "The temperature in <x> is <y> degrees," and sentences from site navigation menus. Additionally, many texts appear to be scraped from adult websites, containing low-quality, machine-translated samples. We have spotted numerous instances of incomplete or significantly incorrect translation pairs. Some target sentences were written in languages other than Ukrainian.

To control the quality of the sentences, we apply multiple heuristics.

Language filtering gcld3 library[1] provides language detection capabilities. We remove all sentences that failed to verify as Ukrainian or English.

Perplexity thresholding We score source and target sentences using two decoder-only models trained on different monolingual datasets (Radford et al., 2019; Minixhofer et al., 2022) and sum their bits per character measures.

Translation mismatch filtering LaBSE (Feng et al., 2022) embeds sentences into a space, where similar sentences in different languages are close together. We use it to filter out badly aligned sentence pairs.

Length filtering The lengths of the original and translated sentences reveal examples that are too short or too long. Absolute differences of lengths point to pairs with long target for the short source and vice versa.

We arbitrarily choose joint values of filtering thresholds to get the desired approximate example counts: 1 million, 3 million and 8 million. We perform multiple experiments with these splits while searching for optimal hyperparameters. We list threshold values in Table 2 and best results for each subset.

4. Second Phase: Unsupervised Data Selection on Extended Multi30K

We use the best checkpoint from the previous fine-tuning phase to train on a high-quality dataset: Extended Multi30K from Saichyshyna et al. (2023). Switching datasets gives us a performance boost of 1.97 BLEU. We additionally delete 11600 sentences from the dataset using unsupervised perplexity filtering pipeline gaining 0.35 on the dev set that translates to 0.3 BLEU on the devtest subset of FLORES.

We use perplexity as a data selection criterion to calculate thresholds to filter out highly surprising sentences. We apply the k-fold cross-validation technique to make the perplexity evaluation in-domain. We split the training data into $k = 5$ folds and train k models withholding one of the folds from each run. Then we score every sentence using the model that has not seen that sentence in training. Next, we sweep for acceptable threshold values by minimizing BLEU on the development set and report results in Table 3. We plot the distribution of scores in Figure 1. We also provide threshold sweep results for training from base Mistral-7B-v0.1 checkpoint in Table 6. By comparing finetuned re-
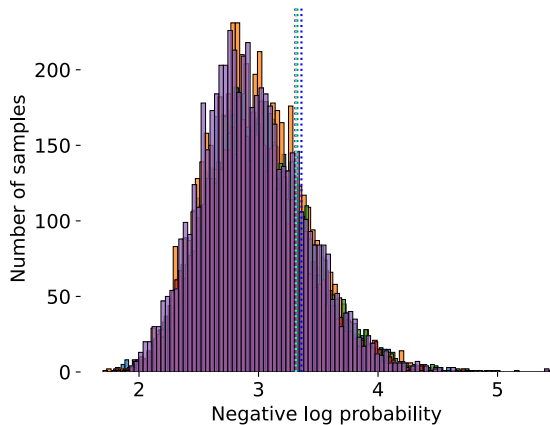
---

[1] https://github.com/google/cld3

Figure 1: Distributions of sentence log probabilities for each fold superimposed on top of each other. Every bar color represents a unique fold; every vertical line denotes a 60th percentile cutoff threshold. The best percentile is chosen using grid search shown in Table 3.

sults, we demonstrate that data from the second phase alone is not enough to match the performance of our best checkpoint.

Threshold percentile	Examples	BLEU ↑ dev	devtest
20th	5800	31.57	32.06
40th	11600	31.65	32.16
50th	14500	31.76	32.36
60th	17400	31.80	**32.34**
70th	20300	31.51	32.17
80th	23200	31.44	32.46
95.4[th] (2σ)	28025	31.74	32.18
Full dataset	29000	31.45	32.04

Table 3: Extended Multi30K log probability thresholds swept on FLORES dev set. We choose the best checkpoint based on model performance on FLORES dev subset using grid search for optimal perplexity threshold value.

5. Few-Shot Translation

Conditioning the model on a sequence of demonstrations of performing some task allows the model to learn this task in-context, also known as "few-shot learning" (e.g. Brown et al. (2020)), thanks to the ability of the Transformers to modulate representations of its future tokens using past context, implementing a specialized internal context-dependent learning algorithm inside its weights (von Oswald et al., 2023).

While few-shot learning allows to quickly try any task with a low number of demonstrations, Liu et al. (2022) have shown that parameter-efficient fine-

tuning allows smaller models achieve better performance, effectively spending less floating point operations per test example at inference time.

Setting up the model for finetuning requires a lot of work, and in-context learning allows to quickly probe capability of a large model using inference software that performs efficient management of key-value cache for speed (Kwon et al., 2023).

To test backbone models before finetuning, we attempt decoding translations with a basic prompt shown in Figure 2.

[INST] They are planning to host a party next weekend. [/INST] Вони планують провести вечірку наступного вікенду.
[INST] I enjoy swimming in the ocean and feeling the salty breeze. [/INST] Мені подобається плавати в океані та відчувати солоний вітер.
[INST]

Figure 2: Basic 2-shot prompt used for few-shot translation. `[INST]` prefixes the beginning of the source sentence and `[/INST]` denotes the beginning of the target translation. These separators are chosen arbitrarily (as in finetuning) and are not special vocabulary items, even though they bear visual resemblance to them.

We find that the model significantly underperforms compared to current state of the art translation models when using beam search (Tillmann and Ney, 2003).

This decoding algorithm performs pruned breadth-first expansion, scoring target sentence prefixes using model's own log probability, approximating maximum a-posteriori estimation of the best translation.

Inspection of the n-best list of translation candidates (beams) reveals that the models can produce high-quality translations, however assign low probabilities to them. We find the best possible translation by rescoring beams using the BLEU score as a loss function (Kumar and Byrne, 2004) with respect to the reference translation (the so-called "oracle").

We employ this oracle rescoring strategy to gauge the potential capability of the model to produce good translations without finetuning, and find that in a regime of increased computation (large width of the beam) and assuming perfect selection capability, a base model is competitive with specialized alternatives. We sweep over a grid of multiple beam widths and report highest attainable BLEU scores in Table 4.

Beams	Oracle BLEU ↑	
	Mistral-7B-v.01	Llama 2 7B
3	27.11	24.55
5	29.20	26.64
10	31.53	28.76
15	32.81	29.09
20	33.54	27.64
25	34.27	26.35
30	33.99	(decoder failure)
35	34.94	
40	34.61	

Table 4: We establish the upper bound of the latent capability of pretrained base models to produce high quality translations with by varying beam width on the task of translating sentences from FLORES dev given a 2-shot prompt. The ground truth oracle determines the best beam. We use beam search implementation by Kwon et al. (2023) with presence penalty of 0.1. The results do not improve monotonically with increasing beam size, and lengths of hypotheses grow with maximum beam size, yielding diminishing returns. This problem can be attributed to label bias (Murray and Chiang, 2018), and rectifying it will require extra regularization.

Consecutive sentences in FLORES are samples from the same document. We hypothesize, dynamically adjusting the prompt by inserting previous translations will improve results. We observe that the model indeed improves translation of certain words such as proper nouns through access to correct definitions provided in the context (Figure 3), however its overall performance degrades in other examples.

We additionally attempt basic 0-shot with a system prompt `You translate English sentences into native Ukrainian.`, and 10-shot prompting using automatic prompt selection based on similarity between source sentences experiments with GPT-4 and GPT-4 Turbo and find that commerical systems perform similarly to other open source systems, as shown in Table 1.

6. Discussion and Limitations

Single-sentence translation Our system is trained on demonstrations of standalone sentence pairs.

Decoder-only models with long context windows We choose to finetune existing decoder-only models since the choice of models with almost the same architecture but different massive pretraining data is abundant. The number of open-source models released recently and their constant improvement offers a good prospective for the machine translation tasks.

Source: RSPCA NEW SOUTH WALES CHIEF INSPECTOR DAVID O'SHANNESSY TOLD THE ABC THAT SURVEILLANCE AND INSPECTIONS OF ABATTOIRS SHOULD BE COMMONPLACE IN AUSTRALIA.

Hypothesis given random 2-shot context: Головний інспектор РСПКА Нового Південного Уельсу Девід О'Шеннесі повідомив ABC, що спостереження та інспекції аббатств повинні бути звичайним явищем в Австралії.

Context example: [INST] ANIMAL LIBERATION AND THE ROYAL SOCIETY FOR THE PREVENTION OF CRUELTY TO ANIMALS (RSPCA) ARE AGAIN CALLING FOR THE MANDATORY INSTALLATION OF CCTV CAMERAS IN ALL AUSTRALIAN ABATTOIRS. [/INST] Організація Звільнення тварин і Королівське товариство із запобігання жорстокому поводженню з тваринами (КТЗЖПТ) знову закликають до обов'язкової установки камер спостереження на всіх австралійських бійнях.

Hypothesis given relevant 2-shot context: Головний інспектор Королівського товариства із запобігання жорстокому поводженню з тваринами (КТЗЖПТ) Нового Південного Уельсу Девід О'Шеннесі заявив, що спостереження та інспекції бійні повинні бути поширеними в Австралії.

Figure 3: Few-shot translation with contextual prompting allows the model to learn named entities on the fly. Without context, the model makes a wrong guess trying to transliterate the abbreviation.

These models receive gradient from all outputs during pretraining, and the self-attention mechanism can see the input, the partial output, and access past examples of translations in its context window using induction heads (Olsson et al., 2022).

For efficiency, we only train on examples with single short sentence pairs and do not pack context windows full of tokens as done in pretraining. In our early experiments, we find that our models still generalize to inputs longer that what is seen in training. This generalization behavior is often attributed to relative position embeddings (Dai et al., 2019; Csordás et al., 2021). We leave evaluation of long context attention stability under these conditions for future work.

Training on the noisy dataset Data cleaning has a positive effect on the resulting metrics. However, our models trained on 8 million filtered, examples perform worse than models trained on 3 million examples (Table 2).

Tokenizer performance We used the LLaMA and Mistral tokenizers during our experiments, which use at least twice as many tokens to compress a sentence in Ukrainian of the same length as an English sentence in character. In practice, that means that generating a sentence in Ukrainian

takes at least twice as many steps to generate. We show a distribution of sentence token lengths in Figure 4.
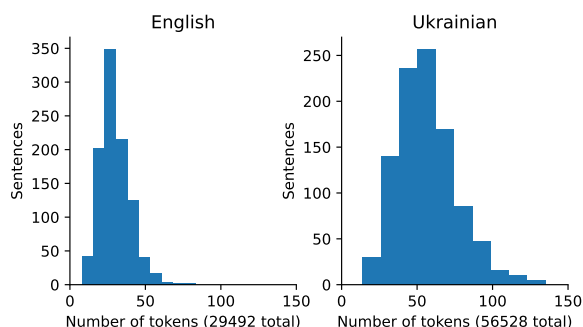

Figure 4: Comparison of tokenizer compression rates between English and Ukrainian using the Mistral-7B tokenizer on the FLORES dev set.

Evaluation   We choose BLEU-4 score (Papineni et al., 2002) as our core evaluation metric and model selection criterion. BLEU-4 measures 4-gram precisions, where grams are defined as words. We use the implementation and rely on tokenization decisions of Post (2018). This metric is sensitive to minor differences that do not affect the meaning of the sentence, for example case inflections that tend to cascade to multiple adjacent words. BLEU is known to poorly correlate with human judgement of translation quality, and Freitag et al. (2022) recommend learned metrics.

Choosing an appropriate learned metric for judgements of translation quality of Ukrainian requires careful consideration, and incorporation of data informed by the language community, such as a curated corpus of grammar corrections that reflects proper modern use of language (Syvokon et al., 2023).

Regardless of limitations of BLEU, improvement in BLEU still signals improvement in translation quality in our regime.

WMT22   Our reviewers have pointed out that WMT22 benchmark (Kocmi et al., 2022) includes a test set for Ukrainian. Our model achieves 24.72 on the WMT22 test set without any postprocessing, ranking behind the best result of Roussis and Papavassiliou (2022) at 25.2 BLEU. We note that the submission that scores relatively low on the WMT22 test, scores comparably to our results on FLORES. These data distribution properties require closer exploration.

7.   Related Work

Translation to Ukrainian   Maksymenko et al. (2023a,b) explore translation controllability by conditioning the model on text embeddings that encode style by finetuning an encoder-decoder model. They claim high quality translations on a private test set.

Instruction-tuned language models   Üstün et al. (2024) explore large-scale translation efforts to produce a multilingual instruction-tuned language model Aya. This work translates large datasets like the Flan Collection (Longpre et al., 2023) using the NLLB-3B model (Team et al., 2022).

Translation systems   Han et al. (2021) provide an iterated backtranslation recipe to bootstrap neural machine translation systems using generative models: zero-shot translation ability is used to produce candidates for few-shot demonstrations. Filtered few-shot demonstrations are used to sample new sentences for further finetuning for translation in two directions.

Translation benchmarks   Besides FLORES-101 (Goyal et al. (2022), or FLORES-200 (Team et al., 2022), both include the same data for Ukrainian) dataset used in this work, Tiedemann (2020) provides an additional dataset for multilingual evaluation.

Data selection techniques   Yang and Li (2023) propose a perplexity filtering pipeline, in which the data is split into k folds to classify low quality augmentation generations produced by surrogate language models. Sachdeva et al. (2024) provide recipes on curating data for language models by directly asking language models to score examples.

8.   Conclusion

In this work, we build a translation system using a two-phase data cleaning pipeline. We demonstrate matching performance to state-of-the-art encoder-decoder models for English-Ukrainian translation task. Notably, our system exhibits superior performance compared to the NLLB model, which was instrumental in generating the Aya dataset and contributed significantly to the advancement of multilingual language models. Improved machine translation could bring new capabilities to the next generation of large language models trained for the Ukrainian language. The recent improvements made for decoder-only backbones and the general dynamics of this process encourages us: we firmly believe that recipes we propose in this paper can be used to improve the quality of the translation by simply upgrading the backbone model.

9. Contributions

Yurii Paniv worked on unsupervised data selection on extended Multi30K dataset, Dmytro Chaplynskyi performed initial training using heuristic filtering of Paracrawl, Nikita Trynus worked on evaluation, Volodymyr Kyrylov designed few-shot learning experiments and evaluation.

10. Acknowledgements

11. References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

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.

Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient fine-tuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever. 2021. Unsupervised neural machine translation with generative language models only. *ArXiv*, abs/2110.05448.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark

Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *arXiv preprint arXiv:2301.13688*.

Daniil Maksymenko, Nataliia Saichyshyna, Oleksii Turuta, Marcin Paprzycki, Maria Ganzha, and Mirela Alhasani. 2023a. Controllability for english-ukrainian machine translation by using style transfer techniques. *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 1059–1068.

Daniil Maksymenko, Olena Turuta, Nataliia Saichyshyna, Maksym Yerokhin, and Oleksii Turuta. 2023b. Controllability for English-Ukrainian machine translation based on specialized corpora. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 1–9, Tampere, Finland. European Association for Machine Translation.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dimitrios Roussis and Vassilis Papavassiliou. 2022. The ARC-NKUA submission for the English-Ukrainian general machine translation shared task at WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 358–365, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

NLLB Team, 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.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Christoph Tillmann and Hermann Ney. 2003. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent.

Heng Yang and Ke Li. 2023. Boosting text augmentation via hybrid instance filtering framework. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1652–1669, Toronto, Canada. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

Appendix

A. Translation Examples

Sample of top 5 worst examples by BLEU from FLORES devtest set   Source: WHILE MOST OF THEIR FOOD WOULD BE FAMILIAR TO US, ROMANS DID HAVE THEIR SHARE OF STRANGE OR UNUSUAL FEAST ITEMS, INCLUDING WILD BOAR, PEACOCK, SNAILS, AND A TYPE OF RODENT CALLED A DORMOUSE

Hypothesis: Хоча більшість їхніх продуктів харчування було б знайомим для нас, римляни мали свою частку дивних або незвичайних страв, включаючи кабана, павича, равликів та гризунів, званих куницями.

Reference: Переважно римляни харчувались тим, що знайоме і нам, але були у них і свої дивні та незвичні бенкетні частування, серед яких дикий кабан, павич, слимаки і гризун під назвою вовчок.

BLEU: 1.72

Source: "I WAS MOVED EVERY TIME WE DID A REHEARSAL ON THIS, FROM THE BOTTOM OF MY HEART."

Hypothesis: "Я був зворушений кожного разу, коли ми робили репетицію з цього приводу, від щирого серця".

Reference: Мене це торкало кожен раз під час репетиції до глибини душі.

BLEU: 1.91

Source: FERAL CHILDREN MAY HAVE EXPERIENCED SEVERE CHILD ABUSE OR TRAUMA BEFORE BEING ABANDONED OR RUNNING AWAY.

Hypothesis: Дитина-дикун, можливо, пережила жорстоке насильство над дітьми або травму, перш ніж бути кинутою або втекти.

Reference: Може таке бути, що дикарські діти були об'єктами жорстокого поводження або отримали серйозні травми перед тим, як їх полишили або вони втекли.
BLEU: 2.31

Source: Workers must often get their superiors' approval for any decisions they make, and are expected to obey their superiors' instructions without question.
Hypothesis: Працівники часто повинні отримувати схвалення своїх начальників на будь-які рішення, які вони приймають, і повинні беззаперечно виконувати вказівки своїх начальників.
Reference: Працівникам часто потрібно отримати схвалення керівника щодо будь-яких рішень, а також від них очікують беззаперечного дотримання настанов керівників.
BLEU: 2.32

Source: Typically there will be a tuition fee to enroll in these educational programs.
Hypothesis: Як правило, буде плата за навчання, щоб зареєструватися в цих освітніх програмах.
Reference: Зазвичай такі освітні програми платні.
BLEU: 2.62

Sample of top 5 best examples by BLEU from FLORES devtest set   Source: The East African Islands are in the Indian Ocean off the eastern coast of Africa.
Hypothesis: Східноафриканські острови знаходяться в Індійському океані біля східного узбережжя Африки.
Reference: Східноафриканські острови знаходяться в Індійському океані біля східного узбережжя Африки.
BLEU: 100.00

Source: Earlier the Chinese news agency Xinhua reported a plane to be hijacked.
Hypothesis: Раніше китайське інформаційне агентство Сіньхуа повідомило про викрадення літака.
Reference: Раніше китайське інформаційне агентство Сіньхуа повідомило про викрадення літака.
BLEU: 100.00

Source: For instance, they didn't have corn, nor tomatoes, nor potatoes, nor cocoa, and no ancient Roman ever tasted a turkey.
Hypothesis: Наприклад, у них не було ні кукурудзи, ні помідорів, ні картоплі, ні какао, і жоден стародавній римлянин ніколи не скуштував індичку.
Reference: Наприклад, у них не було ні кукурудзи, ні помідорів, ні картоплі, ні какао, і жоден стародавній римлянин ніколи не куштував індичку.
BLEU: 90.95

Source: The luminosity and rotation are used together to determine a star's Rossby number, which is related to plasma flow.
Hypothesis: Світність і обертання використовуються разом для визначення числа Россбі зірки, яке пов'язане з потоком плазми.
Reference: Світність і обертання використовуються разом для визначення числа Россбі зірки, яке пов'язане з плазмовим потоком.
BLEU: 83.26

Source: But being placed in the "high tropics" just a few degrees north of equator you will need to deal with both heat (always) and strong sun (when the sky is clear, more rarely).
Hypothesis: Але перебуваючи в "високих тропіках" всього в декількох градусах на північ від екватора, вам доведеться мати справу як з спекою (завжди), так і з сильним сонцем (коли небо чисте, рідше).
Reference: Але, перебуваючи в "високих тропіках" всього в декількох градусах на північ від екватора, вам доведеться мати справу як зі спекою (завжди), так і з палючим сонцем (коли небо чисте, рідше).
BLEU: 82.47

Model	BLEU ↑	spBLEU	chrF	chrF++
Finetuned				
Dragoman P, 10 beams (section 3)	30.38	37.93	59.49	56.41
Dragoman PT, 10 beams (section 4)	**32.34**	**39.93**	**60.72**	**57.82**
Zero shot and few shot (section 5)				
LLaMa-2-7B 2-shot	20.1	26.78	49.22	46.29
RWKV-5-World-7B 0-shot	21.06	26.20	49.46	46.46
gpt-4 10-shot	29.48	37.94	58.37	55.38
gpt-4-turbo-preview 0-shot	30.36	36.75	59.18	56.19
Google Translate 0-shot	25.85	32.49	55.88	52.48
Pretrained				
NLLB 3B, 10 beams	30.46	37.22	58.11	55.32
OPUS-MT, 10 beams	32.2	39.76	60.23	57.38

Table 5: We evaluate generated translations with the sacrebleu library to calculate BLEU, spBLEU, chrF, and chrF++ metrics on the FLORES DEVTEST set. Metric spBLEU was calculated with default BLEU values and tokenizer flores101. Tokenization and detokenization are done using the models' default tokenizers. Evaluation is performed on detokenized sentences with corresponding reference sentences.

Threshold	Examples	BLEU ↑	
percentile		dev	devtest
20th	5800	25.14	25.49
40th	11600	25.39	25.45
50th	14500	25.79	25.93
60th	17400	<u>26.07</u>	**26.01**
70th	20300	26.00	25.72
80th	23200	25.90	26.08
95.4^{th} (2σ)	28025	25.91	25.81
Full dataset	29000	25.74	25.67

Table 6: Evaluation scores for model, finetuned from Mistral-7B-v0.1 directly on Extended Multi30K dataset. We performed log probability thresholds sweep on FLORES dev set. We demonstrate that data from the second phase alone is not enough to match the performance of our best checkpoint. Perplexity filtering improves downstream performance over training on full Extended Multi30K dataset.