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

Problem

- Lack of data for training high-quality Ukrainian/low resource language models.
- Huge part of internet is machine-translated (using bad models!)[1] -> same mistakes occur during model inference.

[1] Thompson, Brian, et al. "A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism." arXiv preprint arXiv:2401.05749 (2024).

Approach

Create a recipe for training SOTA machine translation models with LoRA to expand training datasets with English data.

It consists of two steps:

1. Validate noisy corpus of translation pairs from Paracrawl using different data-validation methods.
2. Continue finetuning on high-quality “Translated Multi-30K” dataset using k-fold perplexity-filtering pipeline.

Target Metric - BLEU

We’ve chosen BLEU score on FLORES dataset for English-Ukrainian translation as a measure of model performance.

First Phase: Heuristic Filtering

We used Paracrawl dataset (13M sentence pairs) and discarded 10M pairs.

Issues identified: repetitive, machine-translated, low-quality samples.

First Phase: Heuristic Filtering

- Language filtering using gcld3 library.
- Perplexity thresholding with monolingual models.
- Translation mismatch filtering using LaBSE distance between source and target sentences.
- Total dataset size selection based on thresholds above.
- Train Mistral-7B-v0.1.

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

Second Phase: Unsupervised Data Selection

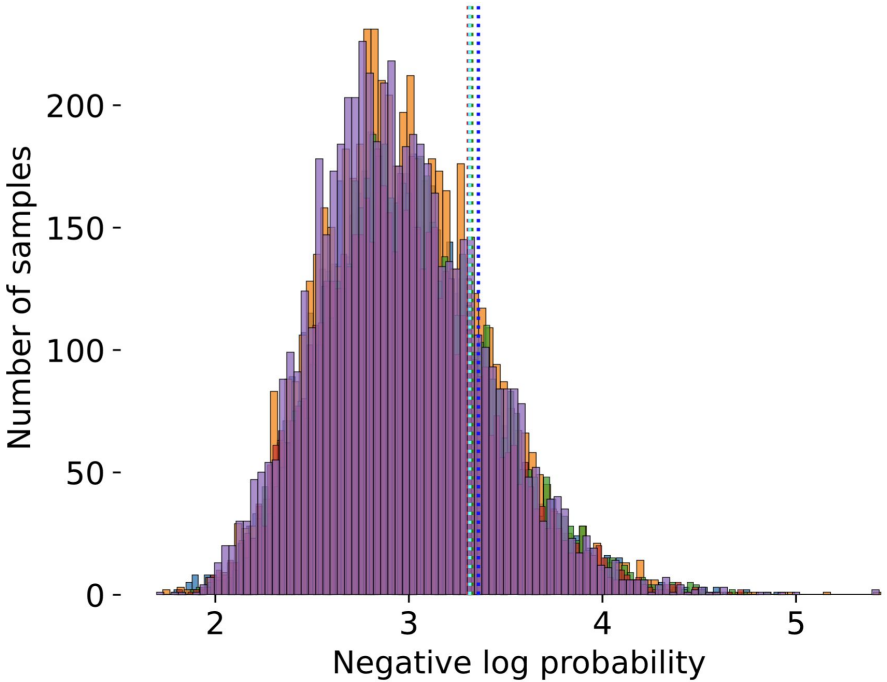
- Continued training on a high-quality Extended Multi30K dataset.
- Used k-fold cross-validation for perplexity evaluation of the dataset.
- Using those scores we applied perplexity-filtering and removed highly surprising sentences by sweeping the perplexity filtering thresholds, **selecting 17K** examples **out of 30K** for final training.

Second Phase: Unsupervised Data Selection

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	<u>31.80</u>	**32.34**
70th	20300	31.51	32.17
80th	23200	31.44	32.46
95.4th (2σ)	28025	31.74	32.18
Full dataset	29000	31.45	32.04

Insight: By **discarding 80%** of the data, we've **achieved nearly the same performance** as with training on full dataset, and even improved the BLEU score by discarding 40% of the data.

Second Phase: Unsupervised Data Selection



Distribution of negative log probability scores for each datapoint in each fold.

Results

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

GPT-4o achieves a BLEU score of 31.4 (improvement over 30.36).

Side effect: translation from other languages

Even though model was finetuned only for English-Ukrainian translation, it translates from other languages as well, even performing GEC task.

- French: “je m’appelle Yura” -> “мене звуть Юра.”
- Spanish: “El español o castellano es una lengua romance procedente del latín hablado” - “Іспанська або кастильська мова - це романська мова, похідна від латинської розмовної мови.”
- GEC: “хто тоимає цей район?” -> “хто володіє цим районом?”

Requires extensive evaluation on benchmarks.

Few-Shot Translation

We investigate resulting beams and we select beam with the highest BLEU score. We found that there are much better translations, but beam-search algorithm couldn’t select them.

Insight: Model already “knows” good translations, but currently there is no algorithm to select them.

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	

Discussion and Limitations

- Our model is finetuned for single-sentence translation.
- Evaluated using BLEU metric on FLORES devtest.
- Our BLEU score on WMT22 benchmark is 24.72 (current SOTA is 25.2 BLEU from Roussis and Papavassiliou (2022)).

We note that submissions that score relatively low on the WMT22 test, scores comparably to our results on FLORES and vice versa, requiring investigation.

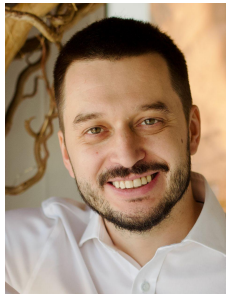
Contributions

- We introduce a recipe to build a translation system using a combination of data cleaning and unsupervised data selection.
  https://github.com/lang-uk/dragoman
- Our model, Dragoman, achieves state-of-the-art performance on FLORES devtest set for English-Ukrainian translation task, achieving **32.34 BLEU**.
  https://huggingface.co/lang-uk/dragoman
- We also released resulting datasets.

Thanks for your attention!

- Online Demo:
  https://huggingface.co/spaces/lang-uk/dragoman

Yurii Paniv



Dmytro Chaplynskyi

Nikita Trynus

Volodymyr Kyrylov