Spivavtor

Співавтор

An Instruction-Tuned
Ukrainian Text Editing Model

UNLP 2024

**Aman Saini, Artem Chernodub,
Vipul Raheja, Vivek Kulkarni**

grammarly

Agenda

1. Introduction
2. Dataset
3. Models
4. Evaluation
5. Learnings

Introduction

What is Spivavtor?

- Spivavtor means “co-author” in Ukrainian.
- It is a collection of open-sourced models trained to perform text editing tasks in Ukrainian.
- Trained on curated, open-sourced Ukrainian text editing dataset.
- Spivavtor supports 4 text editing tasks:
  - Grammatical Error Correction (GEC)
  - Simplification
  - Coherence
  - Paraphrasing
- Spivavtor takes instructions in Ukrainian to perform text editing tasks.
- Spivator models are the Ukrainian equivalent of english-only CoEdit and closely related to mEdit models.

grammarly

What can it do?

GEC Input ▶ Виправте граматику в цьому реченні: А ти, батюшка, стало бути, тут в сторожі? **Output** ▶ А ти, батюшко, стало бути, тут у сторожі?	**GEC Input** ▶ Correct the grammar in this sentence: And you, father, are you here in guard duty? **Output** ▶ And you, father, are you here on guard duty?
Coherence Input ▶ Покращіть зв'язність тексту: Тоді він відвідав острови Греції. Хвороба змусила його повернутися до Риму. **Output** ▶ Потім він відвідав острови Греції, але хвороба змусила його повернутися до Риму.	**Coherence Input** ▶ Improve the coherence of the text: Then he visited the islands of Greece. Illness forced him to return to Rome. **Output** ▶ He then visited the islands of Greece, but illness forced him to return to Rome.
Simplification Input ▶ Спростіть речення: вода спочатку холодна, а потім волога. **Output** ▶ Вода і холодна, і мокра.	**Simplification Input** ▶ Simplify the sentence: first the water is cold, and then it is wet. **Output** ▶ The water is both cold and wet.
Paraphrasing Input ▶ Перефразуйте речення: Наскільки я можу судити, обидва ще живі. **Output** ▶ Наскільки я розумію, вони обидва ще живі.	**Paraphrasing Input** ▶ Rephrase the sentence: As far as I can tell, both are still alive. **Output** ▶ As far as I understand, they are both still alive.

Example model inputs and outputs of the text editing tasks that Spivavtor can perform.
English translations of the examples are provided for reference.

Why did we build it?

Previous work has focused on:

- task-specific Ukrainian text editing like GEC, formality transfer, rather than multi-task text editing.
- general-purpose instruction-tuning (with models like UAlpaca) with no focus on text editing.
- providing massive multi-lingual support for text editing models, but not necessarily focusing on Ukrainian.

Hence, demonstrating a need to build an instruction-tuned model for Ukrainian optimized for text editing.

Dataset

Dataset

- There are many high quality non text-editing Ukrainian datasets[1], and most are either annotated, translated or extracted from multi-lingual datasets.
  - However, there is limited availability of Ukrainian text editing datasets.

- We curated our own dataset called "**Spivavtor-Instruct**", that contains:
  - UA-GEC
    - Publicly available Ukrainian dataset[2] for Grammar and Fluency.
  - Translations from English-only (CoEdIT) datasets using Google Translate API.
    - Simplification - WikiLarge, WikiAuto
    - Coherence - DiscoFuse, IteraTeR
    - Paraphrasing - PAWS

1 - https://github.com/osyvokon/awesome-ukrainian-nlp?tab=readme-ov-file
2 - https://github.com/grammarly/ua-gec

Dataset

- The final dataset also contains task-specific instructions for instruction tuning.
- We prepend task-specific verbalizers in Ukrainian that describe the task to be performed as simple instructions to each instance.
- The Ukrainian instructions were created by native Ukrainian speakers.

Task	Verbalizers	English translation
GEC	"Виправте граматику в цьому реченні:" "Зробіть речення граматичним:" "Удоскональте граматику цього тексту:"	"Correct the grammar in this sentence:" "Make the sentences grammatical:" "Improve the grammar of this text:"
Simplification	"Спростіть речення:" "Зробіть речення простим:" "Зробіть цей текст легше для розуміння:"	"Simplify the sentences:" "Make the sentence simple:" "Make this text easier to understand:"
Coherence	"Виправте зв'язність в реченні:" "Покращіть зв'язність тексту:" "Зробіть текст більш зв'язним:"	"Correct the coherence in the sentence:" "Improve text coherence:" "Make the text more coherent:"
Paraphrasing	"Перефразуйте речення:" "Перефразуйте цей текст:" "Напишіть перефраз для речення:"	"Rephrase the sentence:" "Paraphrase this text:" "Write a paraphrase for the sentence:"

A subset of verbalizers for each task used as instructions in the dataset.

Dataset size

Task	#Training examples	#Validation examples	#Test examples	#Verbalizers
GEC	27,929	3,103	2,682	9
Simplification	11,501	1,278	533	11
Coherence	9,278	1,031	551	7
Paraphrasing	14,076	1,564	6,244	13
Total	62,784	6,976	10,010	40

Models

Model exploration

1 Model architecture

- Encoder-Decoder/ Seq2Seq models
- Decoder-only models
- OpenAI models

2 Model size

- Models with 1B, 7B, 13B parameters.

The selected models were all multilingual with support for Ukrainian.

Instruction tuned models

1 Encoder-Decoder models

- **mT5**
  - google/mt5-large (1.2B)
  - google/mt5-xxl (13B)
- **mT0**
  - bigscience/mt0-large (1.2B)
  - bigscience/mt0-xxl-mt (13B)
- **Aya**
  - cohereForAI/aya-101 (13B)

2 Decoder-only models

- **Bactrian-X**
  - mbzuai/bactrian-x-llama-7b-merged (7B)
- **Mistral**
  - mistralai/Mistral-7B-Instruct-v0.2 (7B)
- **Llama2 chat**
  - meta-llama/Llama-2-7b-chat-hf (7B)
  - meta-llama/Llama-2-13b-chat-hf (13B)

Training

- 8 x A100 GPUs
- AdamW optimizer
- Per-device batch size 8
- Learning rate 5e-5
- Sequence length:
    - 512 for Decoder-only models
    - 256 for Source/Target of Encoder-Decoder models.
- Used Validation cross-entropy loss to pick the best performing checkpoint.

Inference

- Default Generation parameters
- Max output length set to either 512 or 256 depending on the model architecture.
- For Decoder-only models, model-specific EOS tag was used to end decoding.

Evaluation

Baselines

Along with the untrained/base models, we compare against the following baselines in zero shot setting:

- Copy baseline (Target=Source)
- UAlpaca
    - LLaMA 7B model trained on Ukrainian translations of 52K diverse and generic instructions of the Alpaca dataset.
    - Compare the effect of task-specific instruction tuning against large-scale diverse instruction fine-tuning.
- ChatGPT
- GPT4
    - To accommodate for prompt sensitivity, we report the best results among all task verbalizers

Test sets

Task-specific test sets used to evaluate all tasks:

- GEC
  - UA-GEC test set in Ukrainian
- Simplification
  - Asset
  - Turk
- Coherence
  - Discofuse-sports
  - Discofuse-wiki
- Paraphrasing
  - MRPC
  - STSB
  - QQP

Metrics

Evaluation is done for all tasks using the following metrics:

- GEC
  - $F_{0.5}$ Correction score
  - Precision is weighed twice as much as Recall
  - Calculated using ERRANT
- Simplification/Coherence
  - SARI
  - Metric used to evaluate text simplification systems.
  - Calculated using EASSE
- Paraphrasing
  - BLEU
    - Reference-free BLEU / Self-BLEU
    - Reference-based BLEU

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
BACTRIAN-X-7B	D	7B	0.65	36.76	40.37	21.86/8.13
UALPACA-7B	D	7B	0.57	35.17	32.64	13.26/4.95
MISTRAL-7B	D	7B	0.3	38.96	32.41	9.30/3.79
MT0-LARGE	ED	1.2B	0.21	29.56	22.14	6.70/2.68
AYA-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-TURBO	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
SPIVAVTOR-BACTRIAN-X-7B	D	7B	55.73	36.90	47.80	65.31/23.65
SPIVAVTOR-MISTRAL-7B	D	7B	51.54	34.55	44.12	76.56/25.33
SPIVAVTOR-LLAMA2-7B	D	7B	55.88	36.94	48.73	48.97/18.9
SPIVAVTOR-LLAMA2-13B	D	13B	56.48	36.98	48.55	57.31/21.35
SPIVAVTOR-MT5-LARGE	ED	1.2B	61.83	36.40	48.27	77.31/26.68
SPIVAVTOR-MT0-LARGE	ED	1.2B	61.44	36.16	48.28	77.83/26.73
SPIVAVTOR-MT5-XXL	ED	13B	63.00	37.84	48.97	72.42/25.64
SPIVAVTOR-MT0-XXL-MT	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
SPIVAVTOR-AYA-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of SPIVAVTOR models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all SPIVAVTOR models outperform baseline models.

Learnings

Key Takeaways

1

Spivavtor generally performs significantly better over baselines.

Confirming the hypothesis that task specific instruction-tuning results in superior performance on text editing tasks.

2

Domain-specific Instruction tuning outperforms instruction tuning on a large set of generic instructions.

Based on comparisons between UAlpaca and corresponding Spivavtor Llama2 7B model.

3

Encoder-Decoder models outperform Decoder-only models.

For our specific text editing tasks, all Encoder-Decoder models perform better than Decoder-only models.

4

Larger models outperform smaller ones.

Within the same model architecture family, the model performance improves with an increase in model size.

Learning #1:

Spivavtor generally performs significantly better over baselines.

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
BACTRIAN-X-7B	D	7B	0.65	36.76	40.37	21.86/8.13
UALPACA-7B	D	7B	0.57	35.17	32.64	13.26/4.95
MISTRAL-7B	D	7B	0.3	38.96	32.41	9.30/3.79
MT0-LARGE	ED	1.2B	0.21	29.56	22.14	6.70/2.68
AYA-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-TURBO	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
SPIVAVTOR-BACTRIAN-X-7B	D	7B	55.73	36.90	47.80	65.31/23.65
SPIVAVTOR-MISTRAL-7B	D	7B	51.54	34.55	44.12	76.56/25.33
SPIVAVTOR-LLAMA2-7B	D	7B	55.88	36.94	48.73	48.97/18.9
SPIVAVTOR-LLAMA2-13B	D	13B	56.48	36.98	48.55	57.31/21.35
SPIVAVTOR-MT5-LARGE	ED	1.2B	61.83	36.40	48.27	77.31/26.68
SPIVAVTOR-MT0-LARGE	ED	1.2B	61.44	36.16	48.28	77.83/26.73
SPIVAVTOR-MT5-XXL	ED	13B	63.00	37.84	48.97	72.42/25.64
SPIVAVTOR-MT0-XXL-MT	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
SPIVAVTOR-AYA-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of SPIVAVTOR models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all SPIVAVTOR models outperform baseline models.

Learning #2:

Domain specific Instruction tuning outperforms instruction tuning on a large set of generic instructions.

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
BACTRIAN-X-7B	D	7B	0.65	36.76	40.37	21.86/8.13
UALPACA-7B	D	7B	0.57	35.17	32.64	13.26/4.95
MISTRAL-7B	D	7B	0.3	38.96	32.41	9.30/3.79
MT0-LARGE	ED	1.2B	0.21	29.56	22.14	6.70/2.68
AYA-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-TURBO	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
SPIVAVTOR-BACTRIAN-X-7B	D	7B	55.73	36.90	47.80	65.31/23.65
SPIVAVTOR-MISTRAL-7B	D	7B	51.54	34.55	44.12	76.56/25.33
SPIVAVTOR-LLAMA2-7B	D	7B	55.88	36.94	48.73	48.97/18.9
SPIVAVTOR-LLAMA2-13B	D	13B	56.48	36.98	48.55	57.31/21.35
SPIVAVTOR-MT5-LARGE	ED	1.2B	61.83	36.40	48.27	77.31/26.68
SPIVAVTOR-MT0-LARGE	ED	1.2B	61.44	36.16	48.28	77.83/26.73
SPIVAVTOR-MT5-XXL	ED	13B	63.00	37.84	48.97	72.42/25.64
SPIVAVTOR-MT0-XXL-MT	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
SPIVAVTOR-AYA-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of SPIVAVTOR models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all SPIVAVTOR models outperform baseline models.

Learning #3:

Encoder-Decoder models outperform Decoder-only models.

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
Bactrian-X-7B	D	7B	0.65	36.76	40.37	21.86/8.13
UAlpaca-7B	D	7B	0.57	35.17	32.64	13.26/4.95
Mistral-7B	D	7B	0.3	38.96	32.41	9.30/3.79
mt0-large	ED	1.2B	0.21	29.56	22.14	6.70/2.68
aya-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-Turbo	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
Spivavtor-Bactrian-X-7B	D	7B	55.73	36.90	47.80	65.31/23.65
Spivavtor-Mistral-7B	D	7B	51.54	34.55	44.12	76.56/25.33
Spivavtor-Llama2-7B	D	7B	55.88	36.94	48.73	48.97/18.9
Spivavtor-Llama2-13B	D	13B	56.48	36.98	48.55	57.31/21.35
Spivavtor-mt5-large	ED	1.2B	61.83	36.40	48.27	77.31/26.68
Spivavtor-mt0-large	ED	1.2B	61.44	36.16	48.28	77.83/26.73
Spivavtor-mt5-xxl	ED	13B	63.00	37.84	48.97	72.42/25.64
Spivavtor-mt0-xxl-mt	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
Spivavtor-aya-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of Spivavtor models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all Spivavtor models outperform baseline models.

Learning #4:

Larger models outperform smaller ones (within the same model family)

Model	Type	Size	GEC	Simplification	Coherence	Paraphrasing
Copy	-	-	0	21.98	26.89	100/31.4
BACTRIAN-X-7B	D	7B	0.65	36.76	40.37	21.86/8.13
UALPACA-7B	D	7B	0.57	35.17	32.64	13.26/4.95
MISTRAL-7B	D	7B	0.3	38.96	32.41	9.30/3.79
MT0-LARGE	ED	1.2B	0.21	29.56	22.14	6.70/2.68
AYA-101	ED	13B	21.98	35.59	38.30	42.68/15.53
GPT-3.5-TURBO	D	-	1.17	40.18	44.93	26.60/12.51
GPT4	D	-	27.18	40.08	43.44	23.23/11.7
SPIVAVTOR-BACTRIAN-X-7B	D	7B	55.73	36.90	47.80	65.31/23.65
SPIVAVTOR-MISTRAL-7B	D	7B	51.54	34.55	44.12	76.56/25.33
SPIVAVTOR-LLAMA2-7B	D	7B	55.88	36.94	48.73	48.97/18.9
SPIVAVTOR-LLAMA2-13B	D	13B	56.48	36.98	48.55	57.31/21.35
SPIVAVTOR-MT5-LARGE	ED	1.2B	61.83	36.40	48.27	77.31/26.68
SPIVAVTOR-MT0-LARGE	ED	1.2B	61.44	36.16	48.28	77.83/26.73
SPIVAVTOR-MT5-XXL	ED	13B	63.00	37.84	48.97	72.42/25.64
SPIVAVTOR-MT0-XXL-MT	ED	13B	64.55	**38.44**	**49.48**	68.63/25.07
SPIVAVTOR-AYA-101	ED	13B	**64.57**	37.87	48.51	73.28/26.17

Table 3: Comparison of SPIVAVTOR models against various baselines including Copy (target=source), Decoder-only(D) and Encoder-Decoder(ED) models when evaluated in a zero-shot setting. For GEC, we report **$F_{0.5}$ Correction**. For Simplification and Coherence, we report **SARI**. For Paraphrasing, we report **ref-free/ref-based BLEU** where ref-free is the reference-free BLEU and ref-based is the reference-based BLEU to capture the overlap with both source and reference. All scores have been scaled to lie between 0 and 100. Note that all SPIVAVTOR models outperform baseline models.

Task Ablation study

Test model generalization to unseen text editing tasks.

Setting - Hold off one task at a time, train the model on remaining tasks, and measure model performance.

Held-Out Task	GEC	Simplification	Coherence	Paraphrasing
GEC	**18.47**	37.41	52.11	71.44/26.14
Simplification	64.95	**32.84**	48.96	68.39/25.01
Coherence	62.57	36.79	**39.48**	72.86/25.81
Paraphrasing	64.25	36.86	51.84	**74.61/25.90**

Table 4: Performance of the SPIVAVTOR-aya-101 model on all tasks when one task is ablated. We report the same metrics as in Table 3. The bolded numbers represent the zero-shot performance of the model when not trained on that particular task.

Learning - The model generally benefits from seeing task-specific data and has poor performance in a zero-shot setting. The extent to which data helps heavily depends on the task (GEC>Simplification).

Qualitative Evaluation

Qualitative evaluation of the model outputs reveal the following:

1. Baseline models suffer from:
   - Repetitive generation (Decoder only models)
   - Output generation in English
2. OpenAI models suffer from:
   - Task refusal
   - Model admitting no changes are needed
   - Explanation of edits made
3. Spivavtor models correct these mistakes, but aren't perfect either. They suffer from:
   - Excessive truncation in simplification.
   - Replacement of Named Entities with pronouns.
   - Meaning change due to text truncation.

Limitations

Possible limitations of our work:

- Quality of translated datasets depends primarily on the translation API used.
- Scale of the dataset could be improved.
- More metrics around meaning preservation could be introduced.
- Hyper-parameter search is not exhaustive due to time and computational limitations.
- Model performance of API-based closed models could change over time.

Resources

- Our dataset and models are uploaded to Grammarly's Hugging Face collection.
    - Spivavtor dataset
    - Spivavtor models
        - Spivavtor-Large
        - Spivavtor-XXL

- If you have any questions,

  please contact the Spivavtor team.

Link to models and dataset

Thank you

grammarly