UNIWERSYTET
IM. ADAMA MICKIEWICZA
W POZNANIU

UCZELNIA BADAWCZA
INICJATYWA DOSKONAŁOŚCI

EPICUR
EUROPEAN UNIVERSITY

HR EXCELLENCE IN RESEARCH

uam.edu.pl

Wykorzystanie wielkich modeli języka w zadaniu korekty tekstu

Ryszard Staruch

Seminarium ZIL, IPI PAN

Warszawa, 14.04.2025 r.

Zadanie

Korekta błędów językowych w tekstach pisanych.
Przykładowe rodzaje błędów:

- błędy fleksyjne
- błędy interpunkcyjne
- błędy składniowe

Korekta tekstu w języku polskim – stan wiedzy

Zbiory danych

1. The WikEd Error Corpus: Automatic Extraction of Polish Language Errors from Text Edition History (Grundkiewicz, 2013) (https://github.com/snukky/wikiedits)
2. Polish GEC Test Dataset (Ermlab, 2023) (https://github.com/Ermlab/polish-gec-datasets)

WikEd Error Corpus

Zalety:

- rozmiar (miliony edycji)
- teksty pisane przez ludzi
- różnorodne edycje
- otwarta licencja

Wady:

- edycje to nie tylko korekta błędów
- teksty są wyjęte z kontekstu
- brak podziału na train/test

WikEd Error Corpus – przykłady

Saint-simon był raczej bierny **jeslichodzi** o podejście do spraw własnej kariery.
Saint-simon był raczej bierny **jeśli chodzi** o podejście do spraw własnej kariery.

Seattle SuperSonics – klub NBA z Seattle w stanie Waszyngton istniejący od 1967 r. do 2008.
Seattle SuperSonics – **były** klub NBA z Seattle w stanie Waszyngton istniejący od 1967 r. do 2008.

Dziewczyna Wiktor**a(**przyjaciela Adama) pracuje w agencji "Wikam".**Zakochana w Adamie**
Dziewczyna Wiktor**a(**przyjaciela Adama) pracuje w agencji "Wikam".

Ermlab GEC Dataset

Zalety:

- część anotacji ekspertów
- otwarta licencja

Wady:

- większość tekstów bazuje na wikipedii/książkach
- nienaturalne błędy
- brak podziału na train/test

Modele

1. LanguageTool (PL)
2. Modele neuronowe

Prace naukowe

Co jest potrzebne do rozwoju dziedziny w Polsce?

Odpowiedź: Benchmark

Wymagania:

- prawdziwe, różnorodne błędy
- wiedza ekspercka
- wysokiej jakości anotacje

Interdyscyplinarna współpraca CSI & UAM

mgr Ryszard Staruch

prof. UAM Jarosław Liberek

Benchmark – Polish GEC

Cechy benchmarku:

- 304 przykładów dev set
- 500 przykładów test set
- przykłady przygotowane/anotowane przez lingwistów
- teksty pisane przez rodzimych użytkowników języka
- różnorodne domeny tekstów
- autentyczne przykłady
- nowe podejście do ewaluacji

Miary ewaluacji

Standardowa metryka GEC: **F0.5 score**

Na jakim systemie nam zależy?

Odpowiedź: system, którego celem jest znalezienie **wszystkich błędów** w tekście. (recall oriented evaluation)

Wybrana miara: **F1.5 score**

Czy wszystkie błędy powinniśmy traktować tak samo?

Feature name	Dev	Test
Erroneous tokens	5.63%	5.65%
M operation rate	19.98%	18.27%
U operation rate	8.39%	7.63%
R operation rate	71.63%	74.10%
Punctuation error rate	38.18%	35.11%
Letter case error rate	1.65%	1.65%

Propozycja: rozszerzenie Fβ score na **Fβ-raw score**

Fβ-raw score

Jak obliczyć Fβ-raw score?

Fβ score po 2 operacjach:

- lowercasing tekstu źródłowego & poprawionego
- usunięcie znaków interpunkcyjnych

Przykład:
Szanowny Panie, proszę dostarczyć dokument do biura. →
szanowny panie proszę dostarczyć dokument do biura

Ewaluacja wybranych rozwiązań

Parametry ewaluacji LLMów

Temperatura: **0**
Beam size: **1**
Prompt: Dokonaj korekty błędów w następującym tekście zgodnie z zasadami poprawnej polszczyzny. Popraw błędy, ale nie zmieniaj stylu ani płynności wypowiedzi. Zastosuj poprawne symbole właściwe językowi polskiemu, np. cudzysłowy („ ”) oraz myślniki (–). Koniecznie zwróć tylko poprawioną wersję tekstu, bez żadnych dodatkowych wyjaśnień, komentarzy, oznaczeń zmian ani uwag! Tekst do korekty:

Wyniki ewaluacji

Model	Size	F1.5	F1.5-raw
LanguageTool (PL rules)	-	15.38	13.09
Llama-3.1-70B-Instruct	70B	62.61	53.96
Gemma-2-27b-it	27B	**63.17**	56.08
Bielik-11B-v2.3-Instruct	**11B**	53.13	**59.18**
PLLuM-12B-nc-instruct	12B	47.30	50.71
PLLuM-8x7B-nc-instruct	46.7B	47.06	50.91

Różnice między modelami PLLuM

Model	P	R	F1.5	P-raw	R-raw	F1.5-raw
PLLuM-12B-chat	40.36	46.40	44.36	34.90	52.80	45.61
PLLuM-12B-instruct	83.56	35.83	43.47	78.69	39.25	46.41
PLLuM-12B-nc-chat	56.96	48.27	50.65	50.44	53.62	52.60
PLLuM-12B-nc-instruct	89.18	39.14	47.30	86.15	42.87	50.71

- wersje **instruct** proponują bardziej restrykcyjne zmiany
- wersje ***-nc-*** dają lepsze wyniki

Czy zbiór będzie dostępny publicznie?

Zbiór danych będzie dostępny w postaci **wyzwania uczenia maszynowego**.

Adaptacja LLMów do korekty tekstów osób uczących się języka (en)

Problem

- teksty nie brzmią naturalnie
- LLMy wprowadzają zmiany redakcyjne
- **wynikiem** jest wyższy recall kosztem precision
- **celem** są precyzyjne, minimalne poprawy błędów

Jak można dostosować LLMy do zadania?

- Technika data augmentation
- Technika training schedule

Technika – data augmentation

- Powszechna praktyka: odrzucenie przykładów pozbawionych błędów
- Czy ta praktyka ma zastosowanie do LLMów?
- Może warto spróbować odwrotnego podejścia?

Przykład:
Alice **have** a cat. → Alice **has** a cat.

Augmentacja:
Alice **have** a cat. → Alice **has** a cat.
Alice **has** a cat. → Alice **has** a cat.

Data augmentation – wyniki (Gemma 2 9b)

Dataset processing approach	Erroneous sentences		BEA-dev			JFLEG-dev
	FCE-train	BEA-train	P	R	$F_{0.5}$	GLEU
ONLY-ERRONEOUS	100%	100%	60.74	58.79	60.34	58.16
UNCHANGED	65.43%	69.02%	68.99	57.12	66.24	58.73
AUGMENTED	39.55%	40.83%	71.42	53.42	**66.92**	58.21

Technika – training schedule

Pomysł na harmonogram treningu:

1. Trening tylko na błędnych przykładach
2. Trening tylko na poprawnych przykładach z obniżonym learning rate

Intuicja: w drugim etapie możemy kontrolować jak bardzo model nauczy się tego, żeby **nie wprowadzać** zmian w tekście.

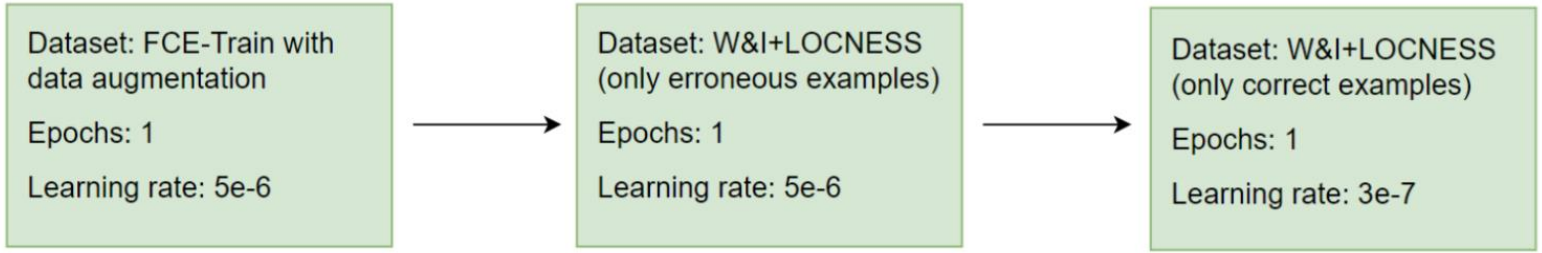

Figure 1: Visualization of the fine-tuning process for our best performing Gemma 2 model on the BEA-dev dataset.

Training schedule – wyniki (Gemma 2 9b)

Learning rate	FCE-train Augmented	BEA-dev			JFLEG-dev
		P	R	$F_{0.5}$	GLEU
1e-7	✗	65.90	58.18	64.19	58.58
1e-7	✓	65.10	58.33	63.62	58.60
2e-7	✗	69.30	56.05	66.17	58.64
2e-7	✓	69.22	56.40	66.21	58.66
2.5e-7	✗	70.94	53.73	66.67	58.47
2.5e-7	✓	70.96	54.40	66.89	58.28
3e-7	✗	73.63	48.72	66.80	57.60
3e-7	✓	73.52	50.10	**67.23**	57.90
3.5e-7	✗	75.81	44.92	66.65	56.74
3.5e-7	✓	75.38	46.82	67.18	57.35
4e-7	✗	77.49	40.15	65.34	55.48
4e-7	✓	76.74	43.49	66.57	56.15
5e-7	✗	79.74	24.79	55.29	50.26
5e-7	✓	78.88	31.78	60.85	52.91

Wyniki na zbiorach testowych

Model	Size	CoNLL-2014-test			BEA-test		
		P	R	$F_{0.5}$	P	R	$F_{0.5}$
T5 Large (Rothe et al., 2021)	700M	-	-	66.04	-	-	72.06
T5 XL (Rothe et al., 2021)	3B	-	-	67.65	-	-	73.92
T5 XXL (Rothe et al., 2021)	11B	-	-	68.75	-	-	75.88
GECToR (Tarnavskyi et al., 2022)	355M	74.40	41.05	64.00	80.70	53.39	73.21
TemplateGEC (Li et al., 2023)	770M	74.80	50.00	68.10	76.80	64.80	74.10
FLAN-T5 XXL (Cao et al., 2023)	11B	75.00	53.80	69.60	78.80	68.50	76.50
DeCoGLM (Li and Wang, 2024)	335M	75.10	49.40	68.00	77.40	64.60	74.40
BART Base (Wang et al., 2024)	400M	76.20	52.20	69.80	77.70	67.50	75.40
Llama-2-13B (Omelianchuk et al., 2024)	13B	77.30	45.60	67.90	74.60	67.80	73.10
Mistral-7B-EPO (Liang et al., 2025)	7B	76.71	52.56	**70.26**	78.16	68.07	75.91
Gemma 2 Augmentation	9B	73.80	**56.16**	69.43	74.86	**71.35**	74.13
Gemma 2 Training-Schedule	9B	75.74	51.47	69.24	79.87	69.90	77.41
Gemma 2 (27B) Training-Schedule	27B	**77.38**	47.88	68.89	**82.28**	67.03	**78.70**

Wnioski

- Istnieje możliwość dostosowania modelu pod wysokie precision
- Operacja w drugą stronę jest cięższa
- LLMy mają obecnie większy potencjał niż modele encoder-only/encoder-decoder

Dziękuję za uwagę!

Zachęcam do zadawania pytań:)

Kontakt:

Adres e-mail: ryszard.staruch@amu.edu.pl
Linkedin: https://www.linkedin.com/in/ryszard-staruch