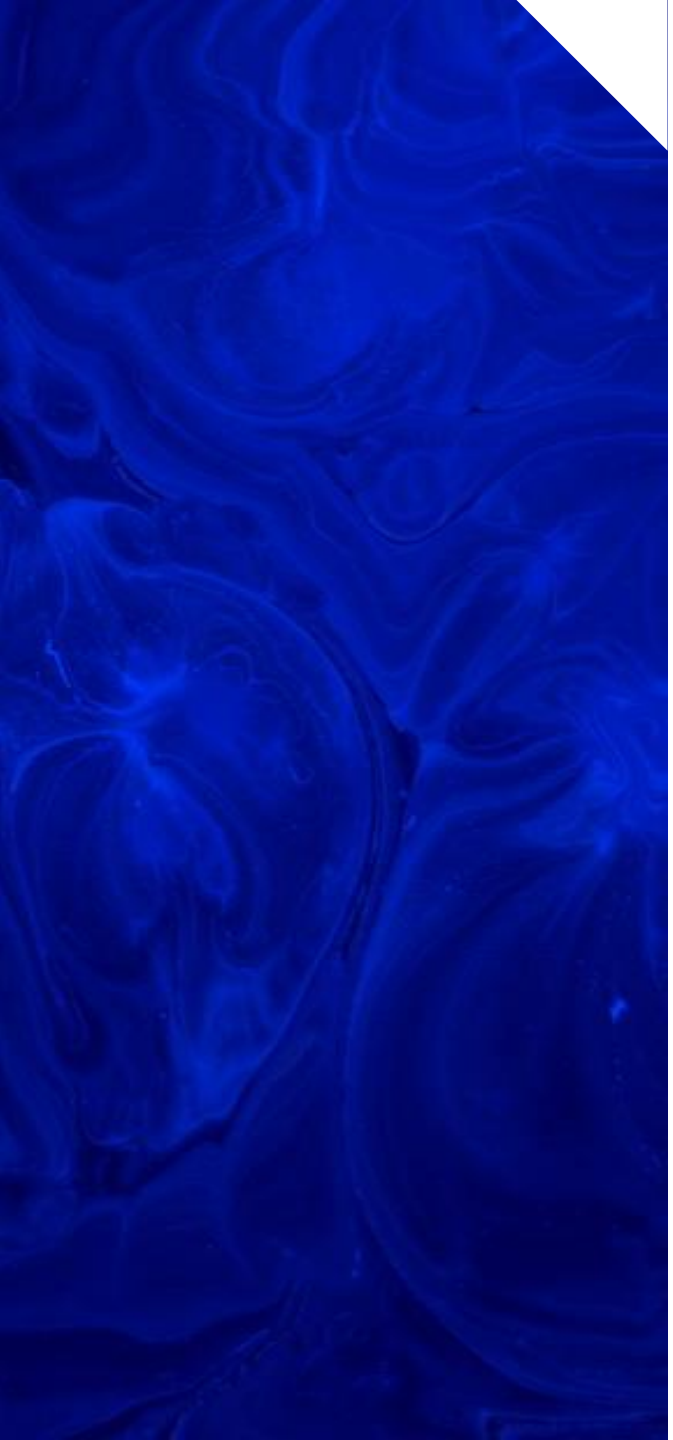


Applica

Simple yet effective method of Post-correcting OCR errors

Paweł Dyda
Adam Mickiewicz University

Competition leader board

Post-correction of OCR results

Overview | Leaderboard | Submissions | Dataset

Submit Solution

☐Contest only	Submitter	Affiliation	Best Submission Description	Task Version	Entries	test-A WER score	test-B WER score
1	Mateusz Piotrowski	None	XXLv1	2.0.1	5	3.725	3.744
2	Krzysztof Wróbel	eNeLPol UJ AGH	ED 3 pl	2.0.1	4		4.302
3	Paweł Dyda	Adam Mickiewicz University & Applica.ai	uml	2.0.1	4	5.338	5.328
4	Zbigniew Bronk	niezależny	v.1	2.0.1	1		7.208
5	Michał Marcińczuk	Samurai Labs, Wrocław University of Science and Technology	Heuristics	2.0.1	10	8.063	8.217

Methodology: problem rephrasing

Idea:

- Use pre-trained Language Model
- Rephrase a problem as Machine Translation task

Models used:

- mt5 (large and base variants)
- plt5 by Allegro (large)

Methodology: data pre-processing

Alignment:

- No alignment
- Full length sequence used

Idea:

- Group books into buckets
- Buckets selected based on publishing year
- Similar number of books per bucket

Methodology: data pre-processing

Table 1: Document counts for a specific bucket

Bucket	Years published	Data counts			
		train	validation	test-A	test-B
19th century	before 1900	7 738	908	1 477	1 406
1900	1900 to 1920	10 294	1 251	1 959	1 992
1920	1920 to 1930	11 310	1 358	2 184	2 066
1930	1930 to 1940	13 525	1 624	2 511	2 420
20th century	post 1940	3 199	366	544	543
other	various	31	1	3	8

Results

Table 2: Results of originally performed experiments

Model	Size	Dataset	WER	
			test-A	test-B
mt5	base	un-escaped	6.305	6.360
mt5	large	escaped	6.504	6.681
mt5	large	un-escaped	**5.338**	**5.328**
plt5	large	un-escaped	5.370	5.378

Errors

\n\n \n\n \n\n \n \n\n\n\n\n\nW ciemnym przysionku czekałem czasami,\nJ
ak na przesmyku, ukryty za drzwiami,\nI pokojówkę, gdy wyszła przypadkiem,\nWiodłem ze sobą, schwyciwszy ukradkiem...
.“\n\n\nPieśń II. w. 159—62.\n\n\n\n\n i t. d. Czyżby nie była można iść do tego gdzieś? — rzekł i rzekł — a może nie może. — A może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a może nie? — rzekł i rzekł — a moż
e nie? — rzekł i rzekł — a nie? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rzekł — a nie może? — rzekł i rze

Errors

```
i DA ya ODRZ TI 0h:\n\nSPIS RZECZY.\n\n    \n\nSZR NOCAdA 2 7 2-1 2 PAZ DW\nI yi Ka : LR
J358\nII 4\n[il 365 ; , 5\nNENA aś2 Zausków (46613 SEAŃ\n\n8\n\nopo *\n\n+\n\nNPPOCA 20682401;
E z - 10\n, VII SZYTE m aSĄAKę PU\n\nVIII RĘT: SZBD 3 =\n\nDOES i a ra AB\n10. X NRZENY 5 >
PULS LUBSKO\nED 43 z 8 19\nRZA. 200 5, . s Ę 38\n\nDZU > 0 4 bać 128\nZY 046 Z 26\nMZZW CL.
7, Gu PE? TAS\n||| AAWLNASZIEJCETENESJEWSENY |.\nBOZY 2 wś ać : .-30\nSENNIK - 030: 1 Sk
« «81\n89. XIX Noki "A\nDRE 0, ż B\n\nrys”\n\n©»\n\n \n          \n\nSPIS RZECZY.\n\n№  Str.*** Jan
Neruda  V\n1. I  3\n2. II  4\n3. III  5\n4. IV  7\n5. V  8\n6. VI  10\n7. VII  11\n8. VIII  13\n9. IX  16\n10. X
17\n11. XI  19\n12. XII  23\n13. XIII  24\n14. XIV  26\n15. XV  28\n16. XVI  29\n17. XVII  30\n18. XVIII
31\n19. XIX  34\n20. XX  35\n\n\n\n SPIS RZECZY. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str. Str.
```

Ablation studies: bucket selection

- No buckets at all
- Clustering by n-grams (k-means over character n-grams)
- Per-year bucket
- Standard

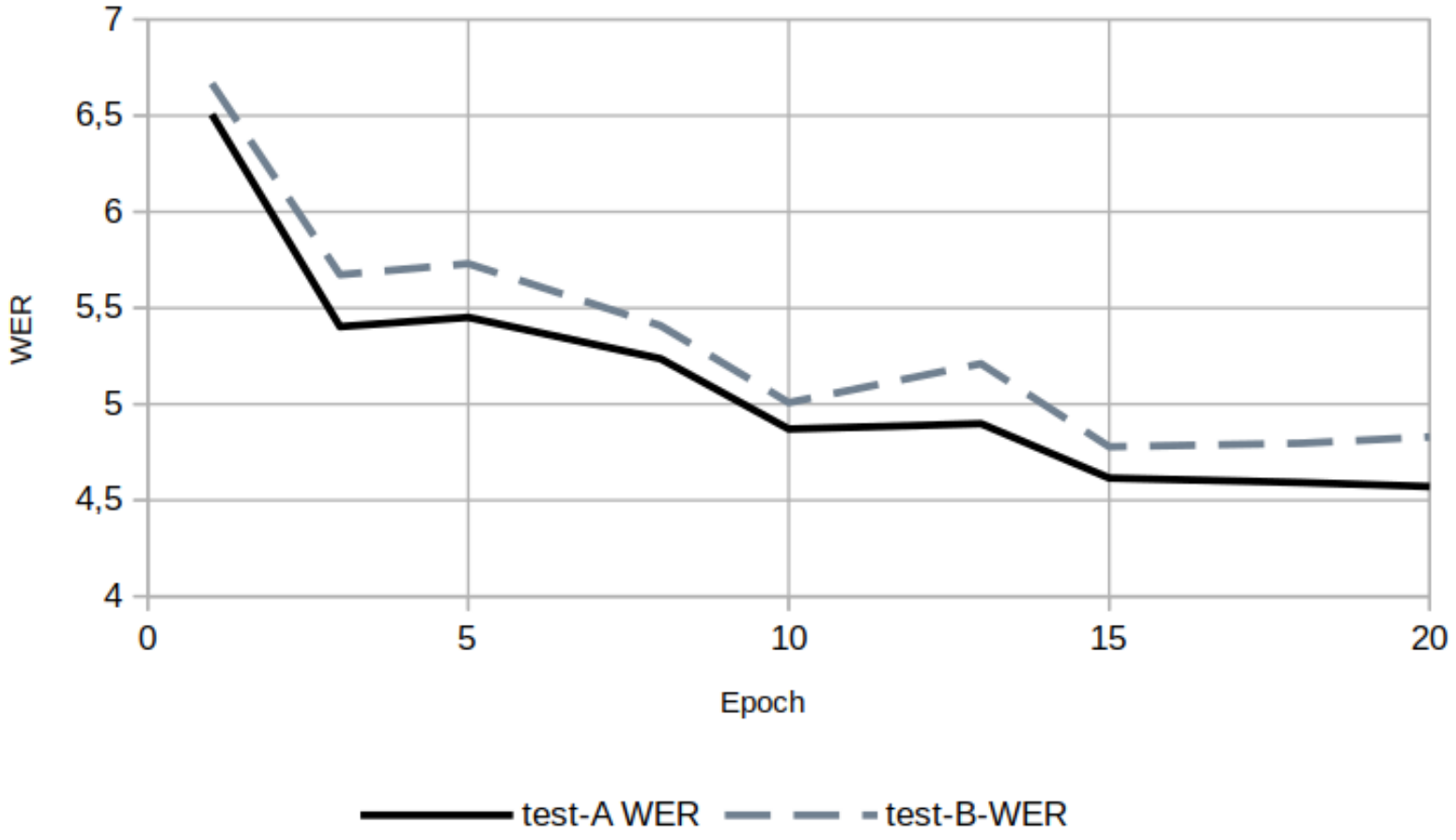
Ablation studies: bucket selection

Table 3: The effect of different bucket assignment method

Method	WER	
	test-A	test-B
standard	5.338	5.328
n-gram clusters	5.157	5.313
no buckets	5.315	5.288
bucket per year	5.152	5.452

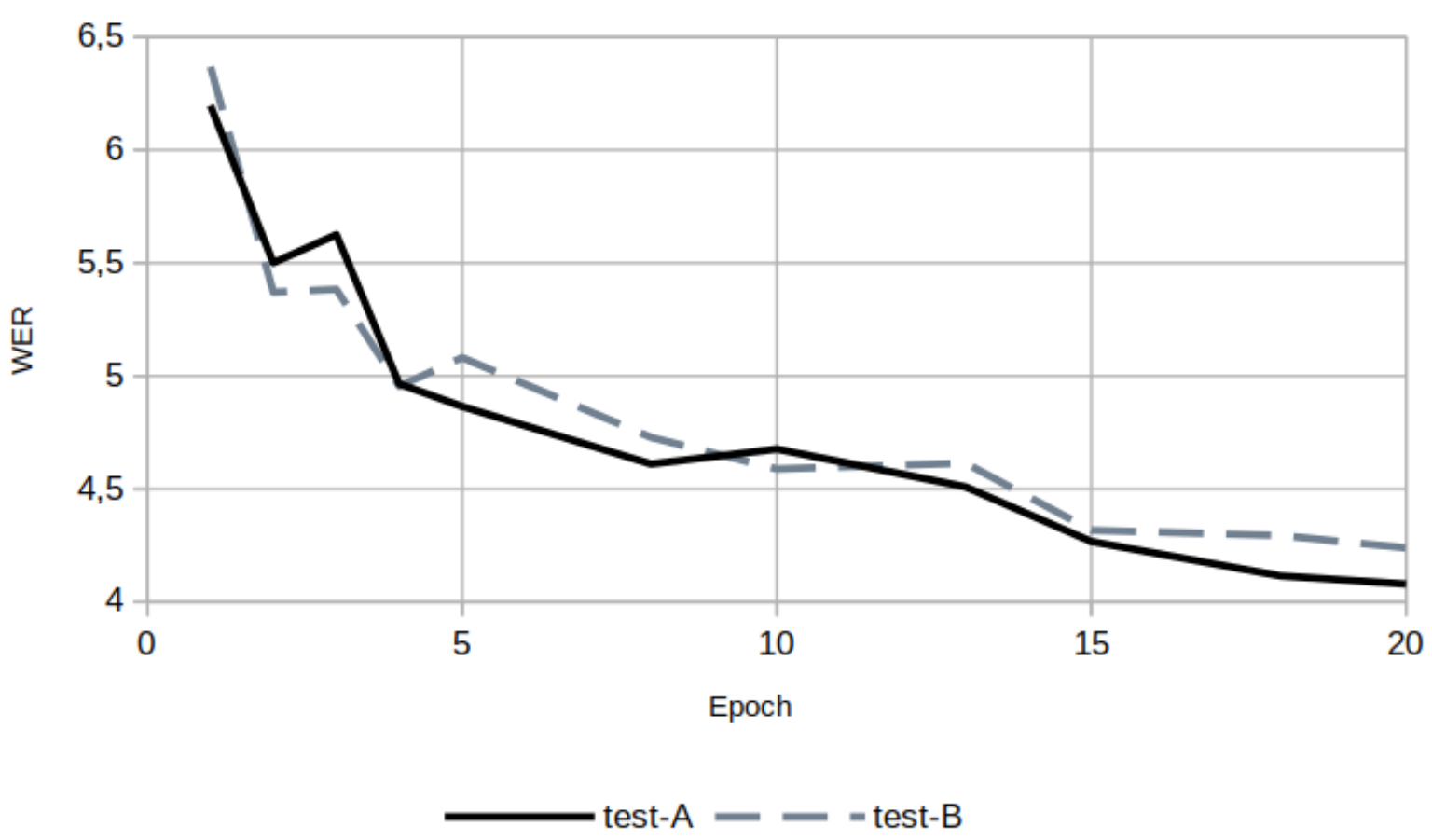
Ablation studies: fine-tuning time

The mt-5 model seem to be under-trained:



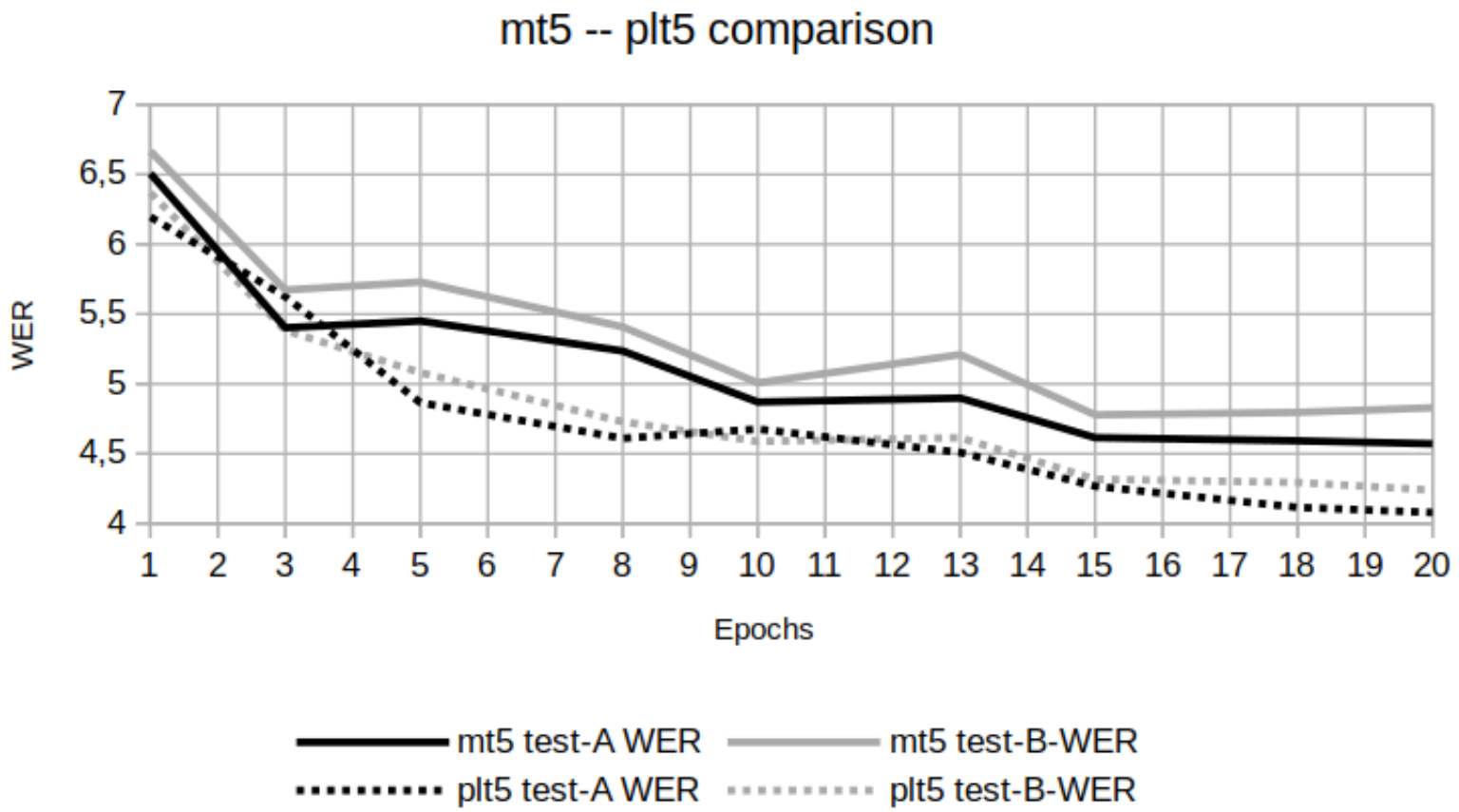
Ablation studies: fine-tuning time

The plt5 model seem also to be under-trained:



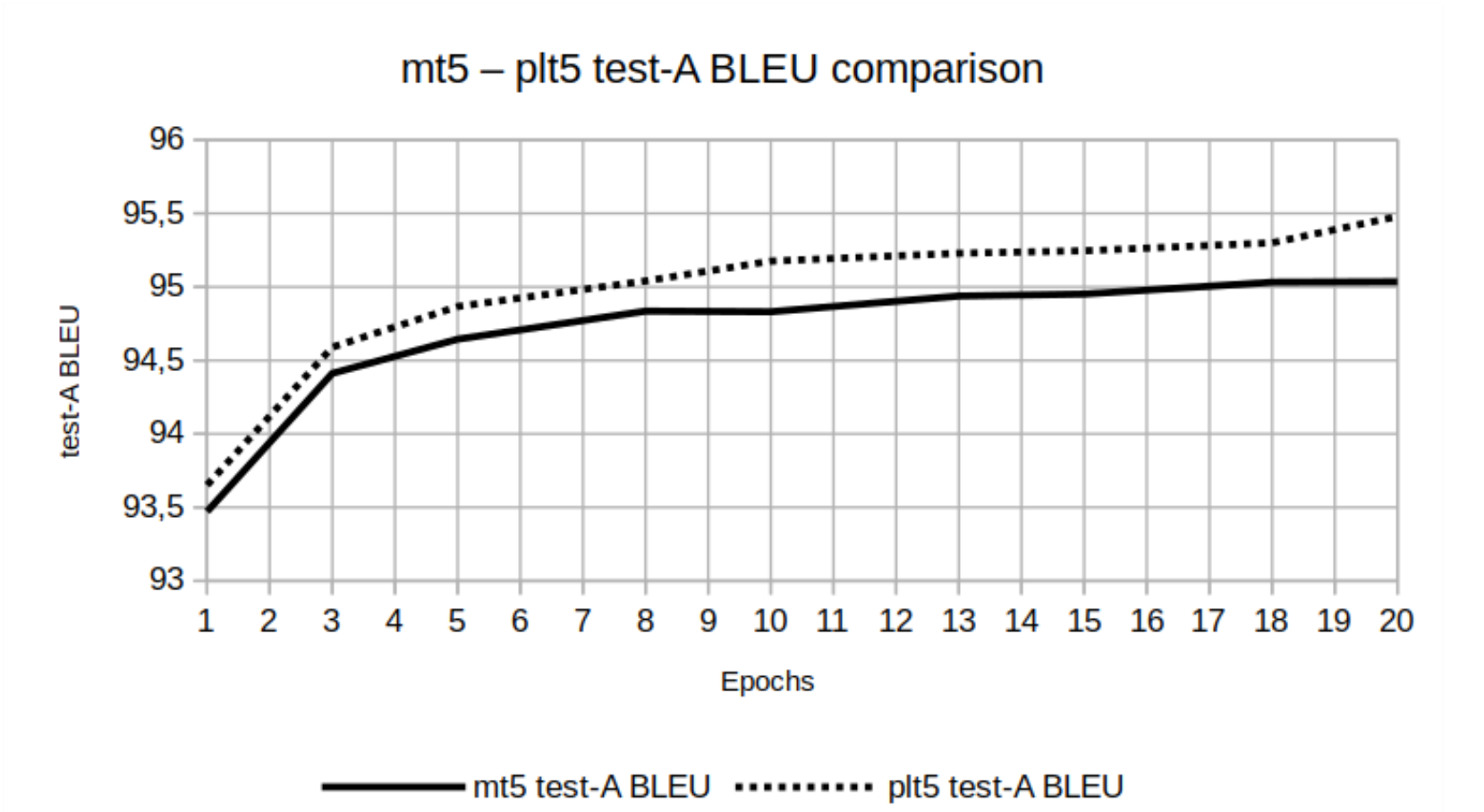
Ablation studies: fine-tuning time

The mt5 – plt5/allegro WER comparison:



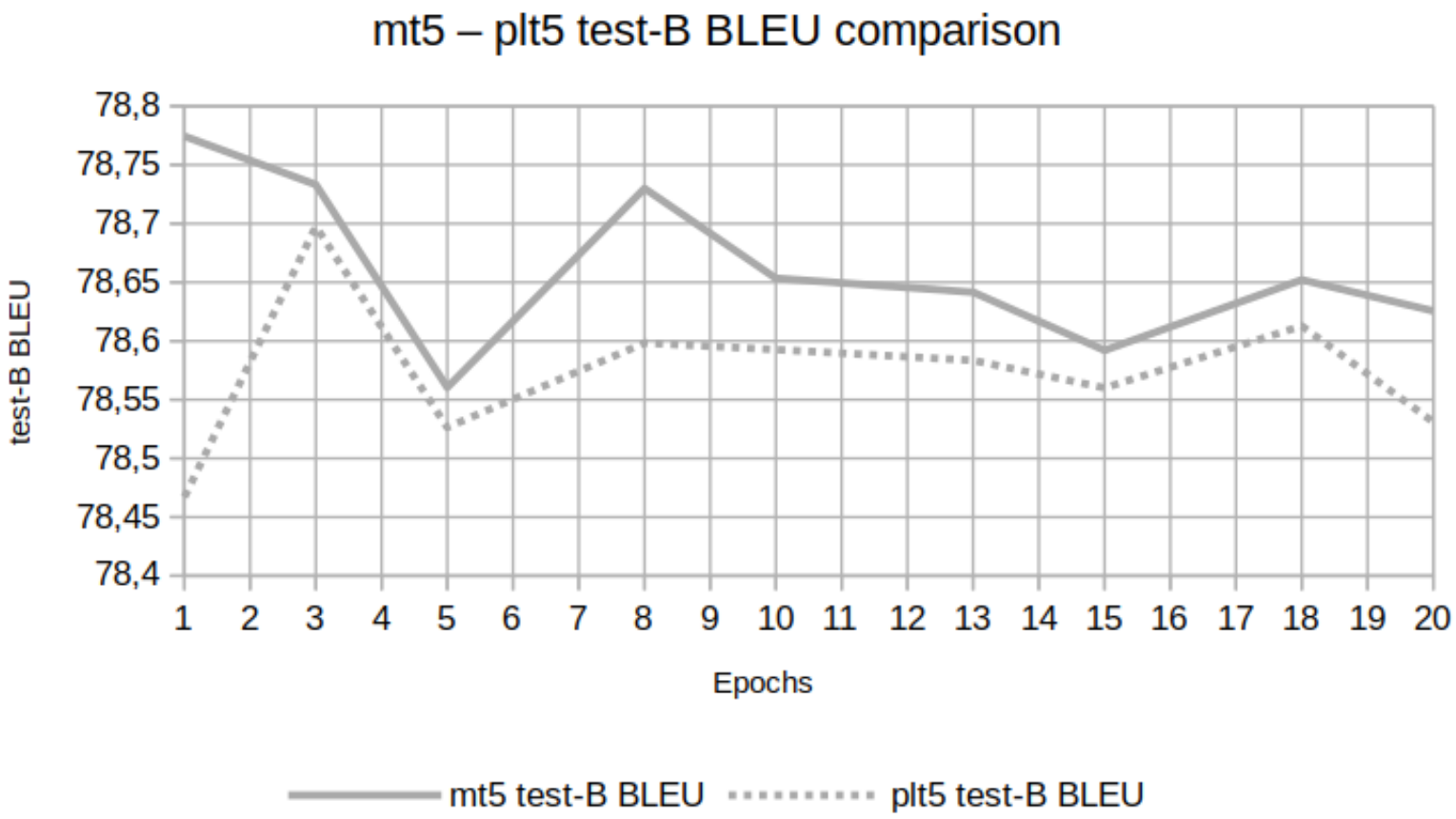
Ablation studies: fine-tuning time

The mt5 – plt5/allegro test-A BLEU comparison:



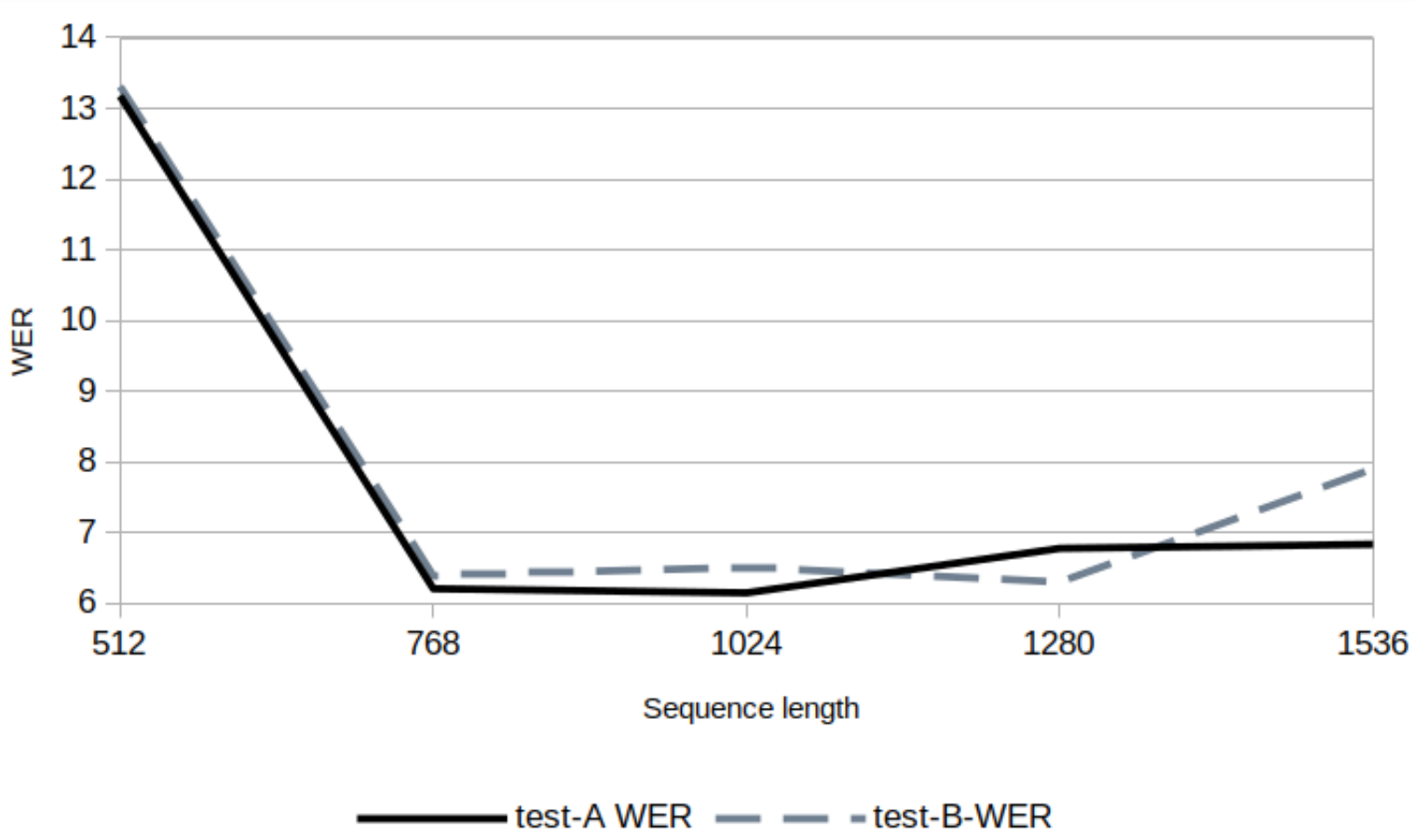
Ablation studies: fine-tuning time

The mt5 – plt5/allegro test-B BLEU comparison:



Ablation studies: sequence length

The mt5-base fine-tuned for 3 epochs:



Thank you

pawel.dyda@applica.ai