How to age BERT well:

Continuous Training for Historical Language Adaptation

Anika Harju & Rob van der Goot

Once upon a time, Old English (≈ 5th–10th century) data was challenging to process. But then, we collected raw data to retrain a BERT model. This led to substantial performance improvements, showing the effectiveness of LM retraining also to a historical language.

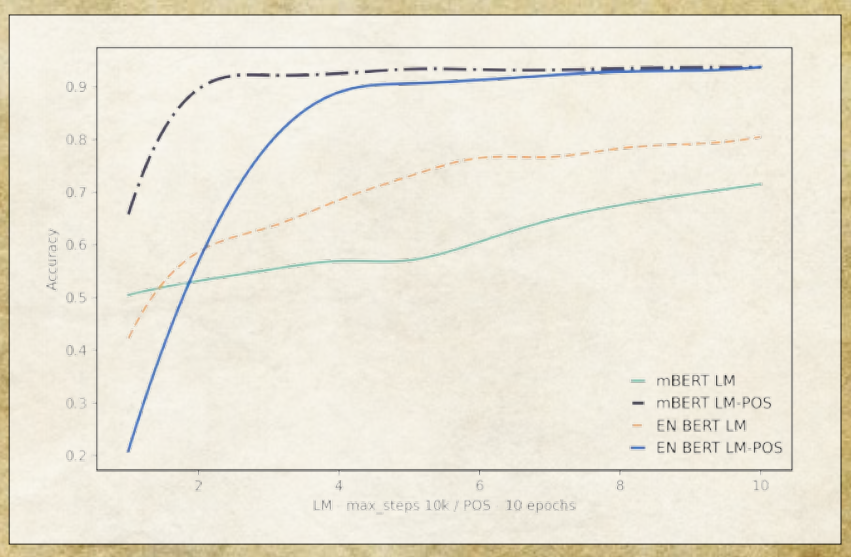
Here will follow an annotated example from the dataset, including a literal translation. First row: original OE data, Second row: literal translation, last row: POS tags:

ac	hi	wunedon	on	clænnysse	oð	heora	lifes	ænde	mid	mycclum	geleafan
but	they	lived	in	purity	until	their	lives	end	with	great	faith
C-	Pp	V-	R-	N	R-	Ps	Nb	Nb	R-	Py	N

Following an extensive data search, we obtained the following:

ISWOC OE treebank	# words	Raw data	# words
West-Saxon Gospels	13,061	Wikipedia	311,793
Anglo-Saxon Chronicles	5,939	Anglo-Saxon Poetry	1,810,636
Apollonius of Tyre	5,541		
Ælfric's Lives of Saints (ood)	3,137		
Orosius (ood)	1,728		

Training on all this obtained data led to the following performance trail:



Different language models challenged each other on multiple battlegrounds, in-domain and out-of-domain:

	In-domain		Out-of-domain	
Model	POS	LM-POS	POS	LM-POS
EN BERT	86.37	92.16	71.96	76.71
mBERT	88.79	93.70	77.87	84.13