A Lithuanian-Specific E5 Text Embedding Model

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

AUTHORS:

Lurda Songailaitė
lurda.songailaite@vdu.lt

Justinas Juozas Dainauskas
justinas.dainauskas@vdu.lt

Milita Songailaitė
milita.songailaite@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

1 INTRODUCTION

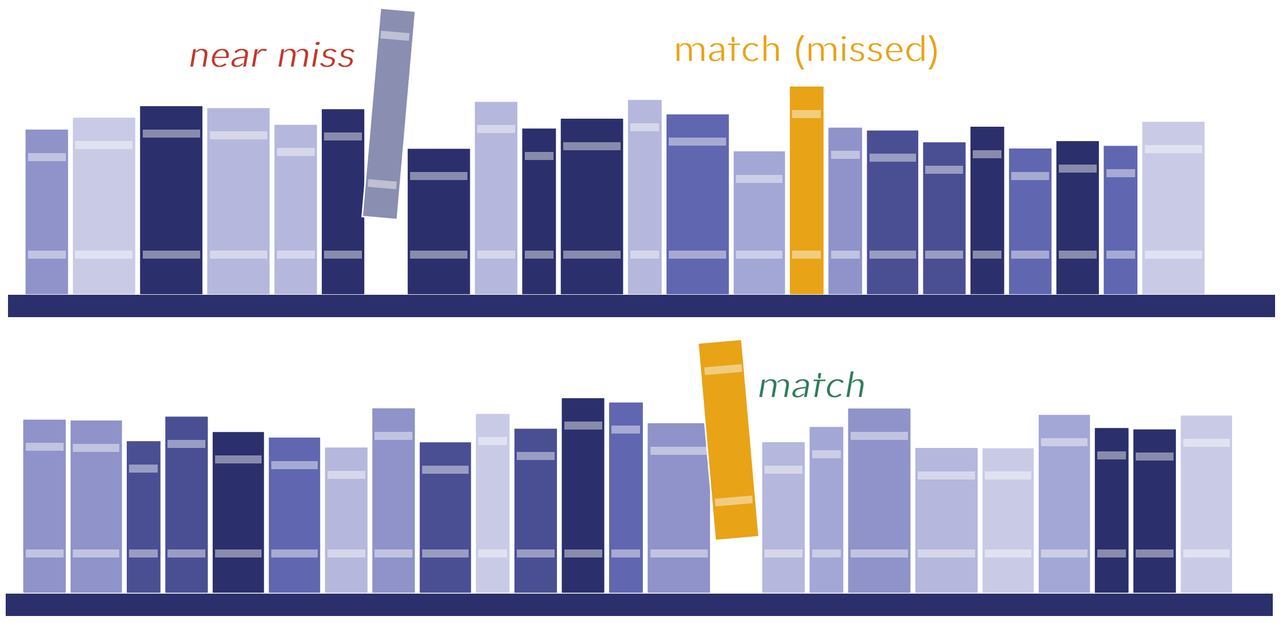
- Lithuanian is an official EU language with 3M+ speakers, yet it remains understudied in NLP.
- No dedicated semantic embedding model; multilingual models see little Lithuanian.
- Rich morphology: words change form with grammatical context, which models trained mainly on English struggle to capture.

2 GOALS

- Train and evaluate an E5 embedding model for Lithuanian.
- Fine-tune multilingual E5 base [1] in two steps.

3 RETRIEVAL TASK

LIBRARY CARD	
Author: *abstract*	
Title: *passage*	
Borrower:	Returned:
Multilingual E5 base	*near miss*
Lithuanian E5 (ours)	*match*



Illustrative: given an abstract, pull its own full text off a shelf of millions

4 METHODS

I Base model

Start from multilingual E5 base [1], a general model trained on dozens of languages at once.

II Weak supervision

Train on a large number of Lithuanian query–passage pairs from multiple datasets, including BLKT [2].

III Hard negatives

Fine-tune on a higher-quality subset of pairs with hard negatives: QA pairs [3] and human-written summaries [4].

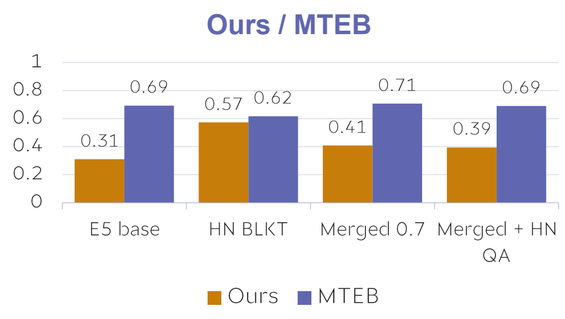
5 TRAINING DATA

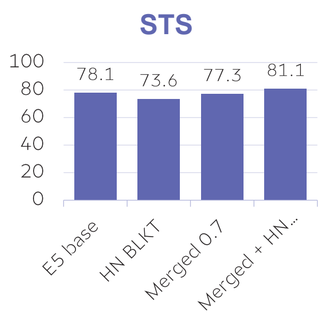
Dataset	Content	Step
BLKT [2]	8.4M texts: news, legal and administrative documents, internet texts, speech transcriptions, fiction	2, eval
QA corpus [3]	10M+ machine-generated QA pairs	2, 3
Summaries [4]	Human-written summaries: IT, law, healthcare, media	3, eval

6 EVALUATION

Lithuanian benchmarks are scarce. We evaluate in three ways: (1) retrieval on a new set built from the summarisation corpus [4], with abstracts as queries and full texts as passages; (2) MTEB tasks; (3) STS, using the English (STS-B) set translated into Lithuanian.

Model	Ours	MTEB	STS
Multilingual E5 base	0.3108	0.6931	78.14
WS on BLKT	0.5733	0.6173	73.61
Merged 0.7	0.4081	0.707	77.27
Merged 0.7 + HN on QA	0.3948	0.6901	81.08





WS = weak supervision, HN = hard negatives. Merged 0.7 = WS model averaged with the base (0.7 tuned + 0.3 base).
Best value per column in amber.

7 CONCLUSIONS

- WS on BLKT raised retrieval on our Lithuanian set from 0.3108 to 0.5733.
- The retrieval gain came with lower MTEB (0.6931 → 0.6173) and STS (78.14 → 73.61).
- Merged 0.7 reached the best MTEB (0.707), and Merged 0.7 + HN on QA the best STS (81.08).

0.3108 → 0.5733
retrieval on our set: E5 base → WS on BLKT

8 FUTURE PLANS

- Extend evaluation with more MTEB tasks and a human-checked Lithuanian STS set.
- Keep improving the model with more data and further training.

[1] Multilingual E5 base model
[2] General Lithuanian Language Corpus (BLKT)
[3] Lithuanian QA corpus
[4] Lithuanian summarisation corpus

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT