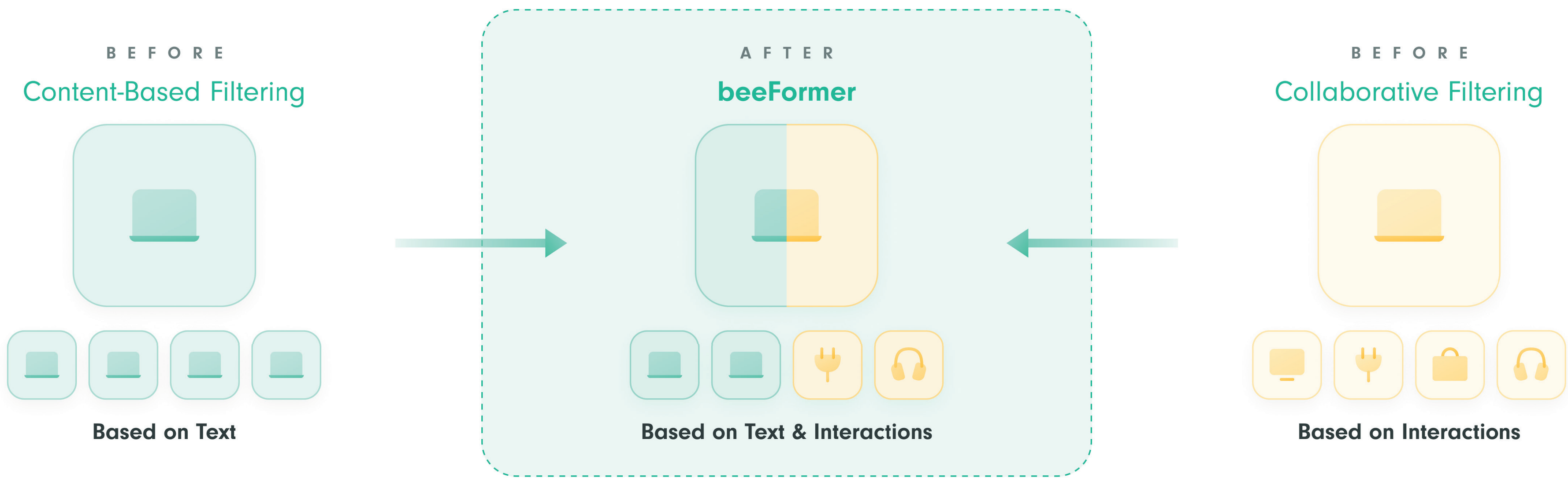
The beeFormer Marks an Important Step Towards Training Domain-Agnostic, Universal Content-Based Models For Recommender Systems

Recombee

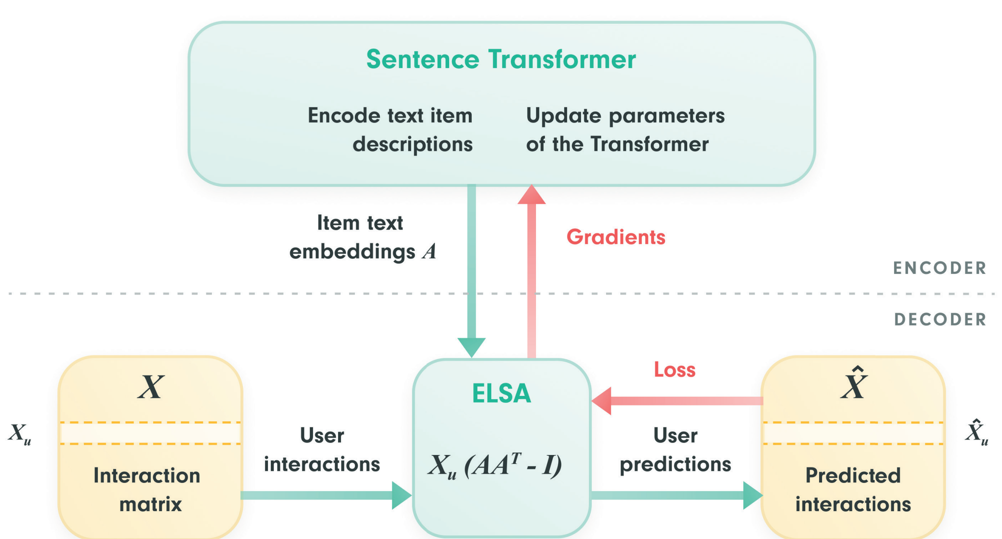
beeFormer: Bridging the Gap Between Semantic and Interaction Similarity in Recommender Systems

Vojtěch Vančura [1,2], **Pavel Kordík** [1,2] and **Milan Straka** [3]

Recombee [1]
Czech Technical University in Prague [2]
Charles University, Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics [3]



Introduction



Collaborative filtering (CF) methods can capture patterns from interaction data that are not obvious at first sight. For example, when buying a printer, users can also buy toners, papers, or cables to connect the printer, and collaborative filtering can take such patterns into account. However, in the cold-start recommendation setup, where new items do not have any interaction at all, collaborative filtering methods cannot be used, and recommender systems are forced to use other approaches, like content-based filtering (CBF). The problem with content-based filtering is that it relies on item attributes, such as text descriptions. In our printer example, semantic similarity-trained language models will put other printers closer than accessories that users might be searching for. Our method is training language models to learn these user behavior patterns from interaction data to transfer that knowledge to previously unseen items.

Method

We train a sentence transformer model in three steps:

- We compute matrix A from text side information without tracking gradients for optimized model
- Than we compute loss (1) with respect to A and create gradient checkpoint
- We optimize weights of sentence transformer model using gradient accumulation

$$L = \left\| \text{norm}\left(X_u\right) - \text{norm}\left(X_u(AA^\top - \mathcal{I})\right)\right\|_F^2 \qquad (1)$$

Results

- Our experiments show that sentence Transformer models trained with beeFormer outperform all baselines in cold-start, zero-shot and time-split recommendation scenarios.
- We demonstrate the beeFormer's ability to transfer knowledge between datasets.
- We show that training models on combined datasets from various domains further increase performance in the domain-agnostic recommendation.
- We create and publish LLM-generated item descriptions for all used datasets for reproducibility of our experiments.
- Models trained with beeFormer are easily deployable into production systems using the sentence Transformers library.

Detailed results: names of our models trained with beeFormer are typed in gray, the best-performing models are represented in bold, and the best baseline for each scenario is underlined.

Zero-Shot Scenario

Dataset	Sentence Transformer	R@20	R@50	N@100
GB10K	all-mpnet-base-v2	0.1017	0.1886	0.1739
	nomic-embed-text-v1.5	0.1146	0.2069	0.1896
	bge-m3	0.1134	0.1953	0.1838
	Llama-movielens-mpnet	0.1782	0.2837	0.2719
	Llama-amazbooks-mpnet	**0.2649**	**0.3957**	**0.3787**
ML20M	all-mpnet-base-v2	0.0788	0.1550	0.1042
	nomic-embed-text-v1.5	0.1113	0.2143	0.1511
	bge-m3	0.1409	0.2125	0.1578
	Llama-goodbooks-mpnet	**0.1589**	**0.2647**	**0.2066**

Cold-Start Scenario

Dataset Method	Sentence Transformer	R@20	R@50	N@100
GB10K CBF	Llama-goodbooks-mpnet	0.2505	0.3839	0.3747
	Llama-goodlens-mpnet	**0.2710**	**0.4218**	**0.4066**
GB10K Heater	all-mpnet-base-v2	0.2078	0.3221	0.3195
	nomic-embed-text-v1.5	0.2154	0.3317	0.3193
	bge-m3	0.2052	0.3113	0.3099
	Llama-movielens-mpnet	0.2060	0.3161	0.3196
ML20M CBF	Llama-movielens-mpnet	0.4291	0.6108	0.4054
	Llama-goodlens-mpnet	**0.4630**	**0.6152**	**0.4066**
ML20M Heater	all-mpnet-base-v2	0.3114	0.4331	0.3407
	nomic-embed-text-v1.5	0.3049	0.4285	0.3270
	bge-m3	0.2847	0.3932	0.3161
	Llama-goodbooks-mpnet	0.3204	0.4669	0.3381

Time-Split Scenario

Dataset Method	Model	R@20	R@50	N@100
zero-shot CBF	all-mpnet-base-v2	0.0218	0.0336	0.0193
	nomic-embed-text-v1.5	0.0387	0.0560	0.0320
	bge-m3	0.0398	0.0546	0.0313
	Llama-goodbooks-mpnet	**0.0649**	**0.0931**	**0.0515**
	Llama-goodlens-mpnet	0.0617	0.0891	0.0492
supervised CF	KNN	0.0370	0.0562	0.0303
	ALS MF	0.0344	0.0580	0.0313
	ELSA	0.0367	0.0628	0.0346
	SANSA	0.0421	0.0678	0.0362
CBF	Llama-amazbooks-mpnet	**0.0706**	**0.1045**	**0.0571**

We want to thank anonymous reviewers for their suggestions, many of which helped us improve our paper. Our research has been supported by the Grant Agency of Czech Technical University (SGS23/210/OHK3/3T/18) and by the Grant Agency of the Czech Republic under the EXPRO program as project "LUSyD" (project No. GX20-16819X).