Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

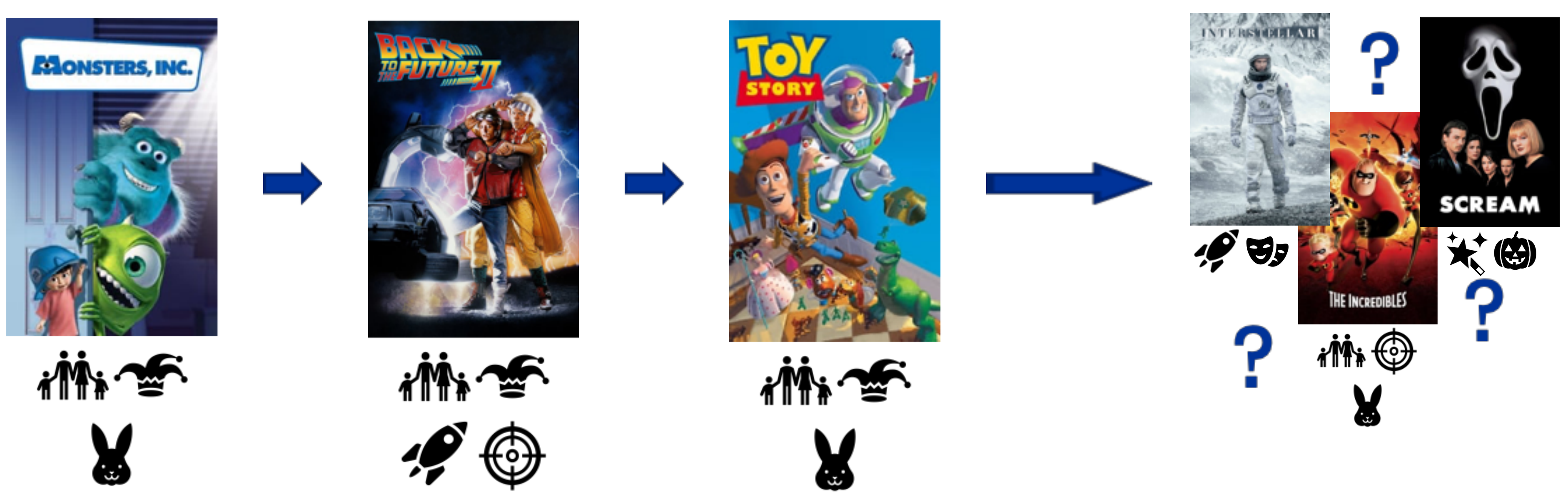
data science

Integrating Keywords into BERT4Rec for Sequential Recommendation

Elisabeth Fischer, Daniel Zoller, Alexander Dallmann, Andreas Hotho
{elisabeth.fischer, zoller, dallmann, hotho}@informatik.uni-wuerzburg.de

BERT4Rec[1] is a recent adaption of the BERT[2] model from the NLP domain for the task of sequential recommendation and has proven itself to be state-of-the-art. One limitation is the representation of items merely as ids, as there is often additional information about items available. Therefore we propose **KeBERT4Rec**, a modification which allows to **add keyword descriptions** for each item.

Which movie should we recommend?

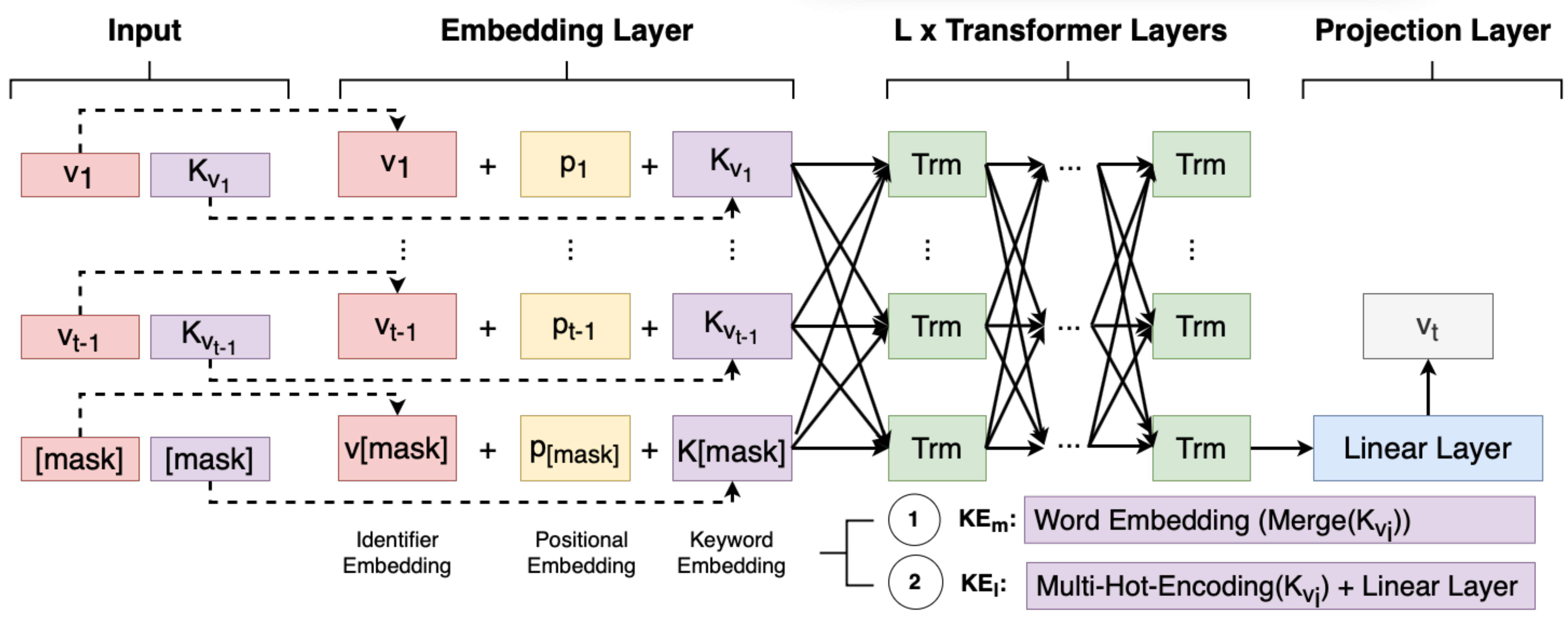


Knowing that all movies in the sequence are comedy and family movies, the horror movie "Scream" is unlikely to be the next one. "Interstellar" and "The Incredibles" share some of the genres already seen and are therefore more likely to be selected.

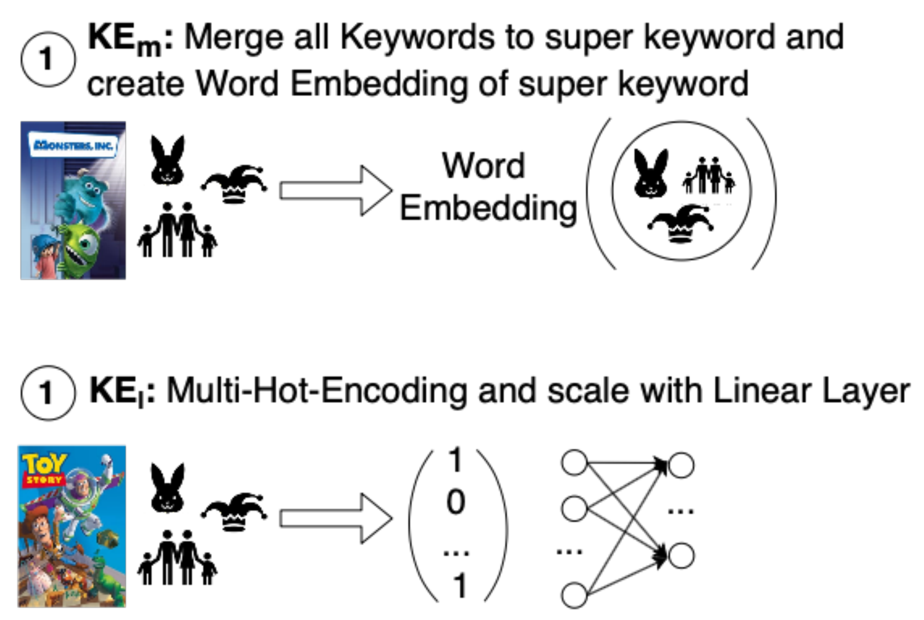
BERT4Rec with Item Keywords

- given a sequence of items BERT4Rec is able to predict the next item
- BERT4Rec is operating on item ids and their positions in the sequence
- additional information about the items is often available and might be useful for recommendation
- we integrate such information in the form of keywords into the BERT4Rec model to improve the predictions
- for each item we use a set of keywords as the item description
- we propose two variants to embed these keywords into our architecture

KeBERT4Rec Architecture



Example: Keyword Embeddings



Datasets

Movielens 20m: ML-20m contains movie ratings from MovieLens. Genre information of movies is used as keywords.
Fashion: Click paths collected from an online fashion shop, with keywords describing the content.

dataset	#User	#Items	#Keywords	#Interactions	Avg.length	Density
ML-20m	138,493	26,744	20	20m	144,4	0,54 %
Fashion	47,158	63,706	301	1.2m	24.4	0.02 %

Results

- popular baseline is outperformed by both BERT4Rec and KeBERT4Rec
- LastItem baseline (only available for Fashion) is also outperformed by both methods
- both variants with keyword embeddings are significantly better than BERT4Rec
- bigger performance increase observed on Fashion dataset, which might be explained by more granular keywords
- multi-hot keyword encoding (KE_l) performs significantly better in most cases and never worse than KE_m, so we can conclude it is the better approach.

Dataset	Metric	POP	Bert4Rec	KE_m	KE_l
ML-20m	HR@1	0.022	0.528	0.536	**0.542***
	HR@5	0.081	0.871	0.876	**0.877**$^+$
	HR@10	0.138	0.943	**0.946**	0.945
	NDCG@5	0.051	0.715	0.722	**0.725***
	NDCG@10	0.070	0.739	0.745	**0.747***
Fashion (LI: 0.294)	HR@1	0.029	0.476	0.642	**0.648**$^+$
	HR@5	0.066	0.700	**0.824**	0.823
	HR@10	0.089	0.795	0.871	0.871
	NDCG@5	0.048	0.594	0.741	**0.743***
	NDCG@10	0.056	0.625	0.757	**0.759**$^+$

Results of experiments as HitRate (HR) and Normalized Discounted Cumulative Gain (NDCG). Both variants of KeBERT4Rec are significantly better than BERT4Rec ($\alpha \leq 0.01$). KE_l marked with * is significant better than KE_m with $\alpha \leq 0.01$ and $^+$ with $\alpha \leq$ 0.05.

1. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W. & Jiang, P. (2019). BERT4Rec. Proceedings of the 28th ACM International Conference on Information and Knowledge Management - CIKM
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding