Towards Interpretable Embeddings: Aligning Representations with Semantic Aspects

Nitisha Jain, Antoine Domingues, Adwait Baokar, Albert Meroño Peñuela, and Elena Simperl

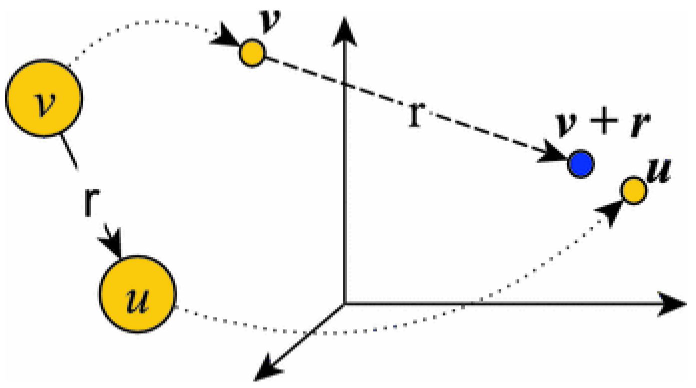
29 November 2024

"Bringing Back Semantics to Knowledge Graph Embeddings: An Interpretability Approach." *International Conference on Neural-Symbolic Learning and Reasoning (NeSy) 2024.*

Knowledge Graph Embeddings

- Embed components of KG (entities, relations) into continuous vector spaces.
- Allow easy manipulation of data while preserving inherent structure of KG.
- Several popular KGE models - TransE, RESCAL, DistMult, ComplEx, ConvE ..
- Many applications - KG completion, rule-based reasoning, entity clustering, relation similarity etc.

KG triple <v, r, u >



Translation based KG embedding

Interpretability Limitations of KGEs

- The dimensions of learned vector spaces do not normally correspond to semantically meaningful properties.
- This limits the interpretability of learned vector space representations.
- Similar entities are not clustered together in the vector space (Jain et al.)
- Previous work (Derrac et al.) on mitigating this issue - identify interpretable directions in learned vector spaces that can play the role of quality dimensions.

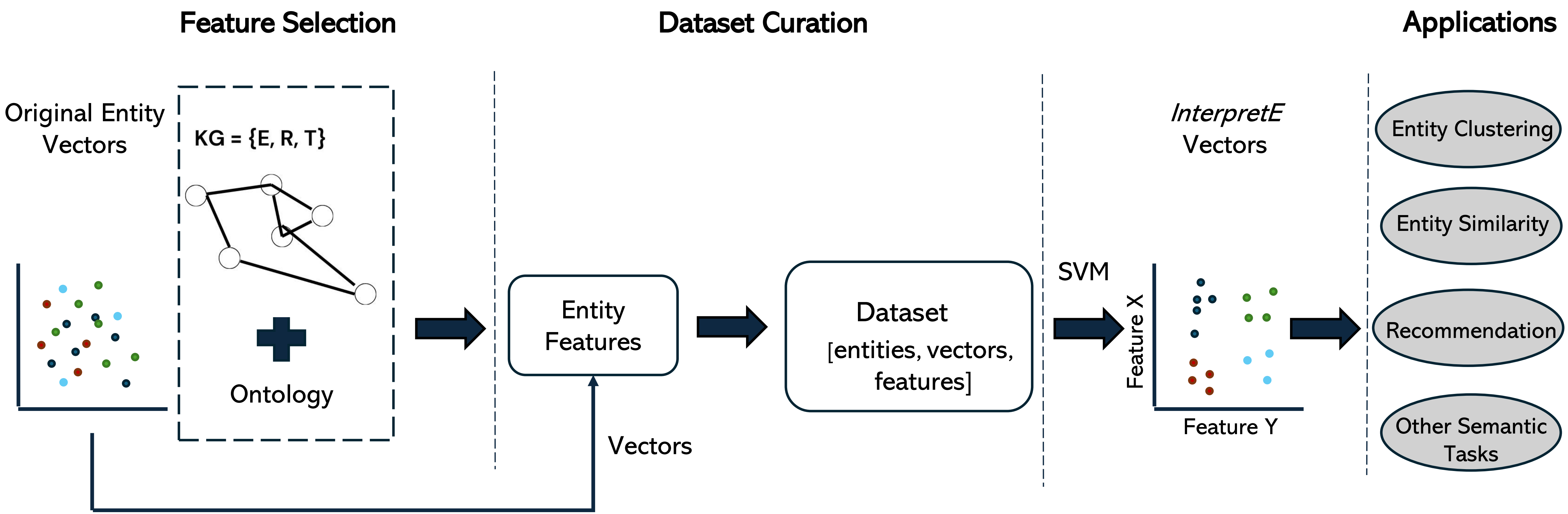
Nitisha Jain, Jan-Christoph Kalo, Wolf-Tilo Balke, Ralf Krestel: Do Embeddings Actually Capture Knowledge Graph Semantics?. Proceedings of the Extended Semantic Web Conference (ESWC), 2021.

*****Joaquín Derrac and Steven Schockaert. Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. Artif. Intell., 66–94, 2015.***

Proposal - *InterpretE*

- *InterpretE* is a novel neuro-symbolic approach to derive interpretable embeddings (from any KG embedding model) for the KG entities.
- *InterpretE* embeddings encapsulate the desired semantic aspects of the entities.
- The method is highly flexible in terms of the number and types of aspects that it can work with, making it scalable for different datasets.
- Desired user-selected or task-oriented entity aspects are identified and selected from underlying datasets through a data-driven process.

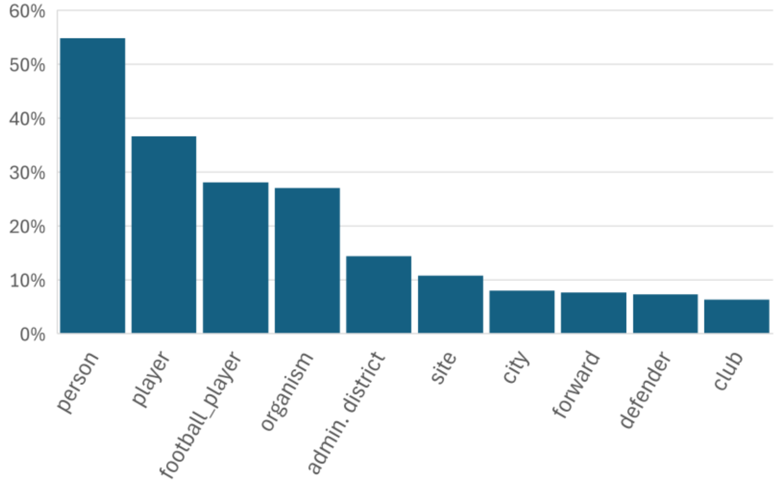
Overview of *InterpretE*



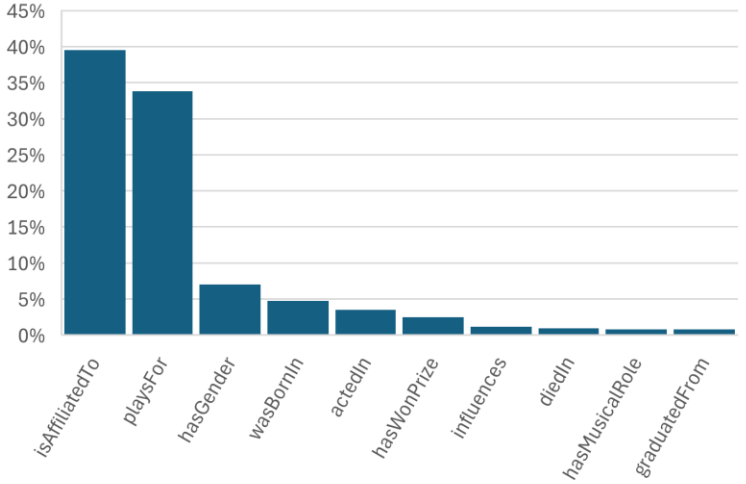
Data Analysis and Selection of Entity Aspects

- Yago3-10 and FB15k-237
- Wordnet-based entity type mappings (Jain et al. 2021), most frequent types and relations chosen.
- Next, most represented, prominent values for a given relation extracted.
- These values, coupled with the associated relation, serve as the entity aspects or features.
- Entities were labeled with binary values indicating the presence or absence of each aspect.

Top 10 most represented entity classes in YAGO3-10



Top 10 most represented relations for *person* entities in YAGO3-10



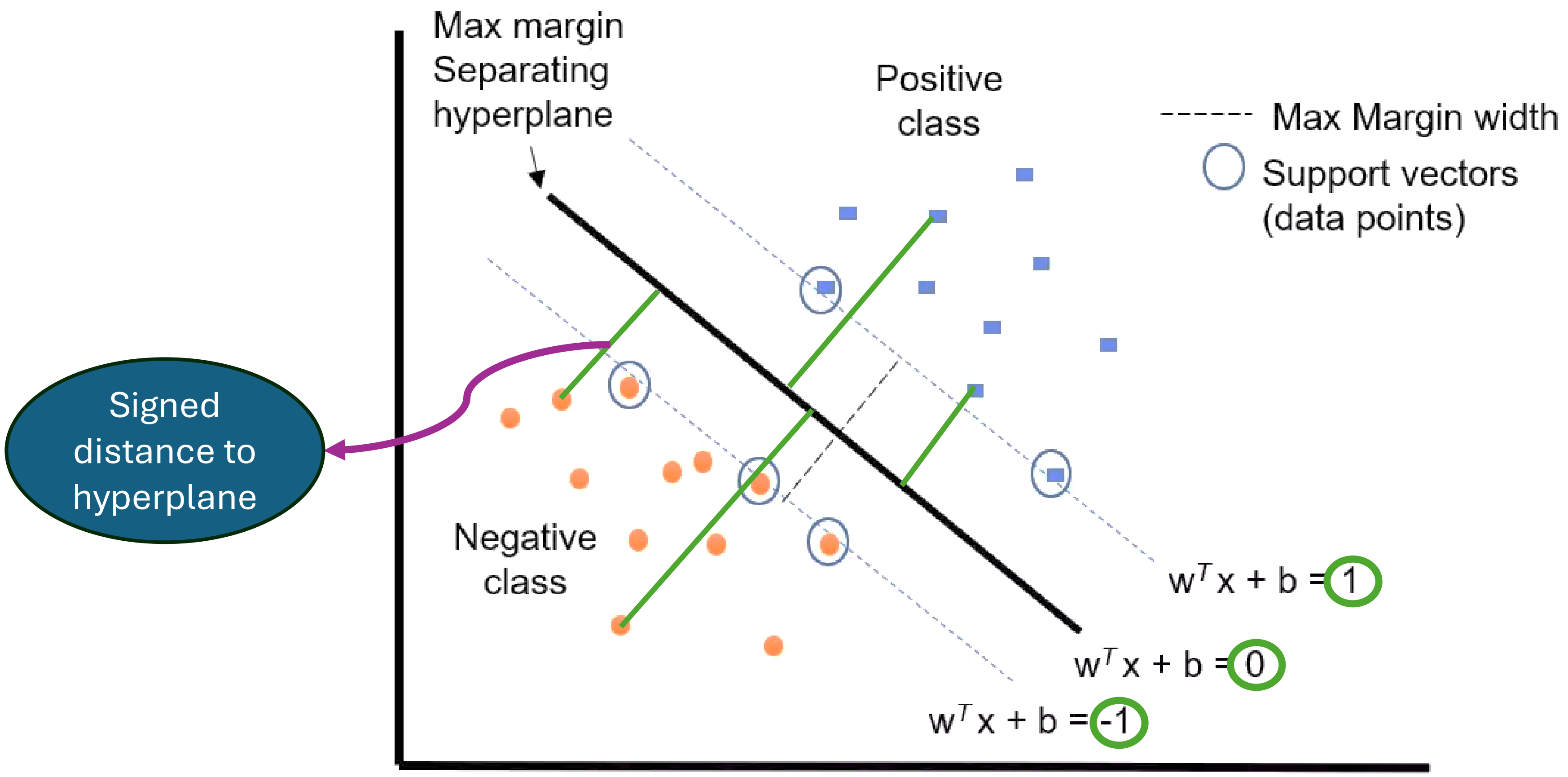
Abstraction of Features

Organization+ isLocatedIn		
	total	6652
Country	#	%
United States	2341	35,19%
United Kingdom	458	6,89%
Canada	414	6,22%
Japan	392	5,89%
France	261	3,92%
Australia	186	2,80%
Unknown	144	2,16%
Germany	144	2,16%
Italy	131	1,97%
India	121	1,82%

- Different levels of abstraction were considered for the features of the entities.
- E.g., for *organization* entities, the relation *islocatedIn* was significant.
- Mapping done for locations - cities grouped by their respective countries or continents.
- This allowed for evaluations across varying abstraction levels.

Deriving *InterpretE* with SVM classifiers

- SVM classifiers trained for each feature separately (based on Derrac et al.)
- Pre-trained KG embedding vectors mapped to a new interpretable vector space.

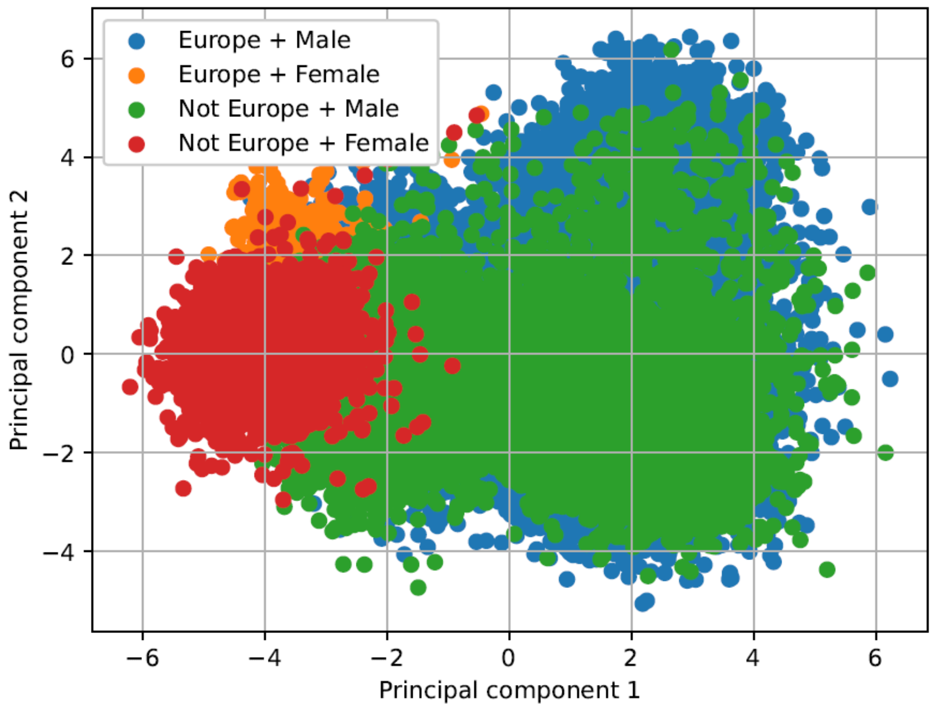


Decision function as a coordinate for the current feature

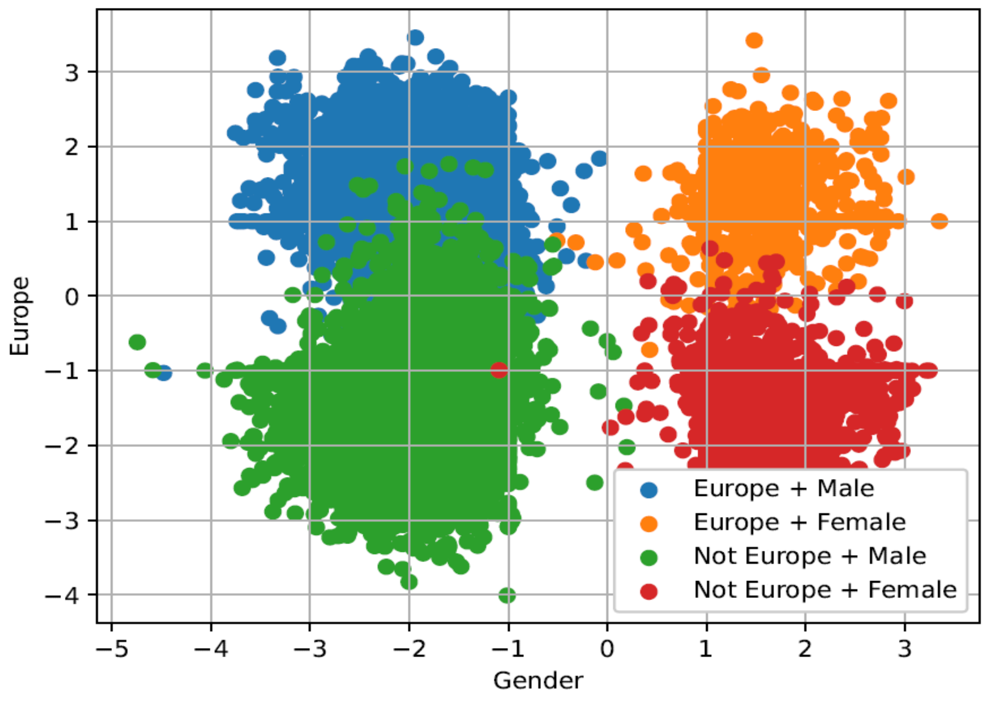
Learnt dimensions in the vector space correspond to human-understandable features of the entities.

Original KGE vs *InterpretE*

***ComplEx* vectors**

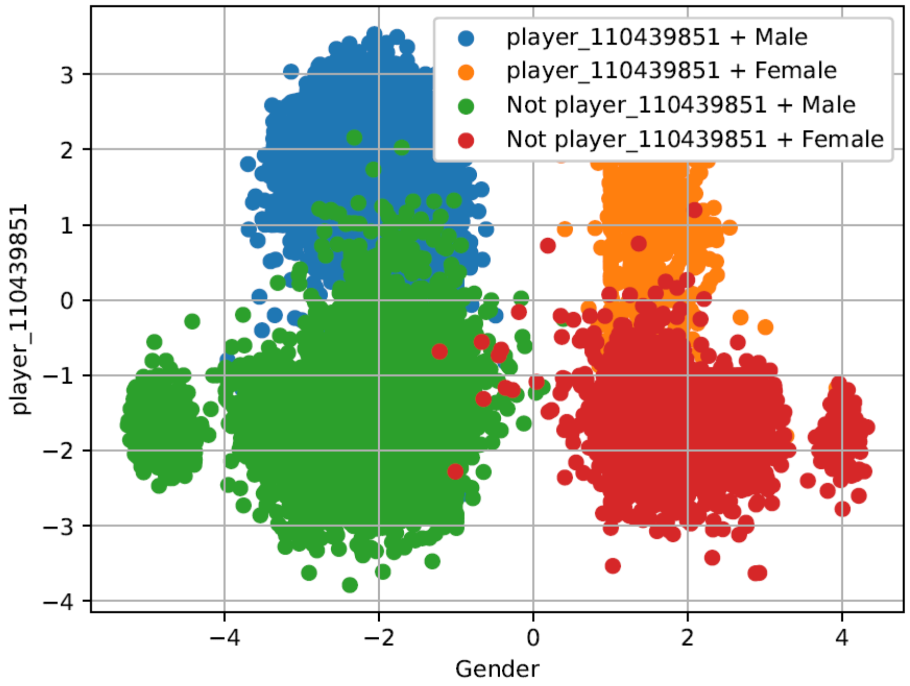


***InterpretE* vectors**

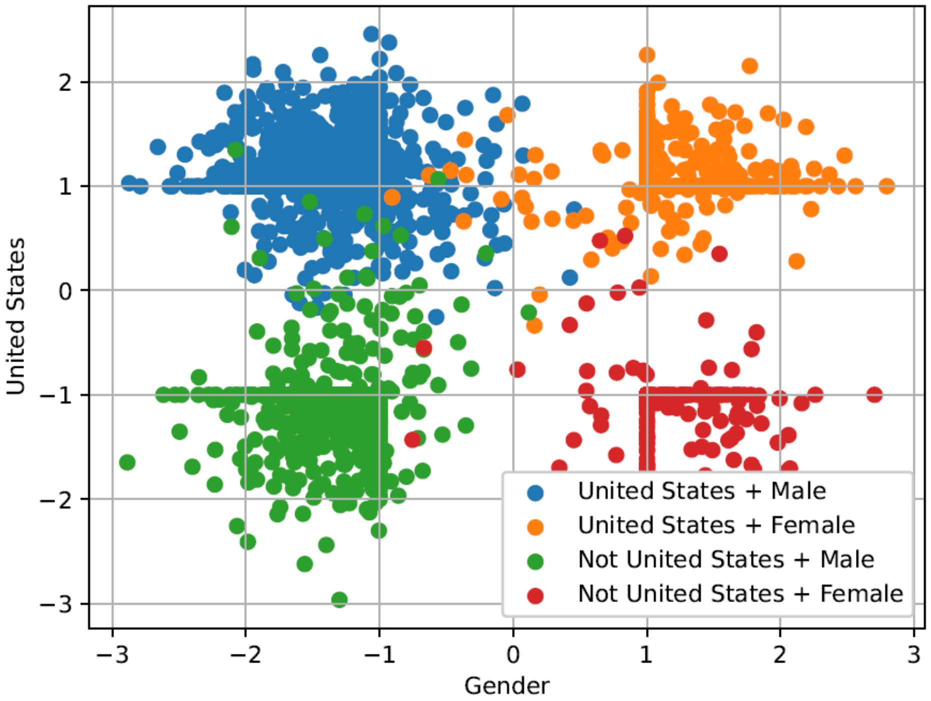


2D projection of vectors for class person and features *hasGender* and *wasBornIn* “Europe" in Yago3-10

InterpretE Examples



InterpretE vectors for class *player* and feature *hasGender* in Yago3-10



InterpretE vectors for class *person* and features *hasgender* and *place_of_birth* "United States" in FB15k-237

Evaluation of *InterpretE*

SVM Evaluation - *Cohen's Kappa Score*

- Evaluates the level of agreement between two sets of categorical labels - the predictions made by the trained SVM and the ground truth labels for the test entities.
- The κ score ranges from -1 to 1, with values closer to 1 indicating a stronger alignment.

Value of κ	Strength of Agreement
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
> 0.80	Very Good

Semantic Evaluation – *Simtopk*

- *InterpretE* vector spaces yield entity vectors organized into clusters aligned with the selected features.
- Evaluate the semantic similarity of the derived vectors (in terms of the features) by measuring the similarity of entities' neighbors.

$$simtopk = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k}\sum_{j\in N_i(k)} f(n_i, n_j)\right)$$

n : the number of total entities; k : the number of considered neighbours;
Ni(k) : the k closest neighbours of the i-th entity, determined using a euclidean distance;
f (· , ·) : returns 1 if the two entities are similar in terms of features, 0 otherwise.

Results

κ scores on the test set and *simtop10* scores on the original and *InterpretE* vectors (mean values) with FB15K-237 for the different KGEMs

Number of features		ConvE	TransE	DistMult	Rescal	Complex
1	κ score	.90	.80	.90	.90	.85
	original	.211	.210	.214	.215	.210
	InterpretE	**.322**	**.298**	**.313**	**.322**	**.319**
2	κ score	.89	.8	.9	.9	.89
	original	.336	.329	.342	.343	.335
	InterpretE	**.484**	**.480**	**.493**	**.514**	**.509**
5	κ score	.72	.68	.72	.65	.73
	original	.561	.538	.545	.523	.547
	InterpretE	**.853**	**.844**	**.889**	**.882**	**.868**
6	κ score	.84	.73	.83	.88	.84
	original	.587	.524	.575	.563	.563
	InterpretE	**.952**	**.918**	**.936**	**.956**	**.932**

Results

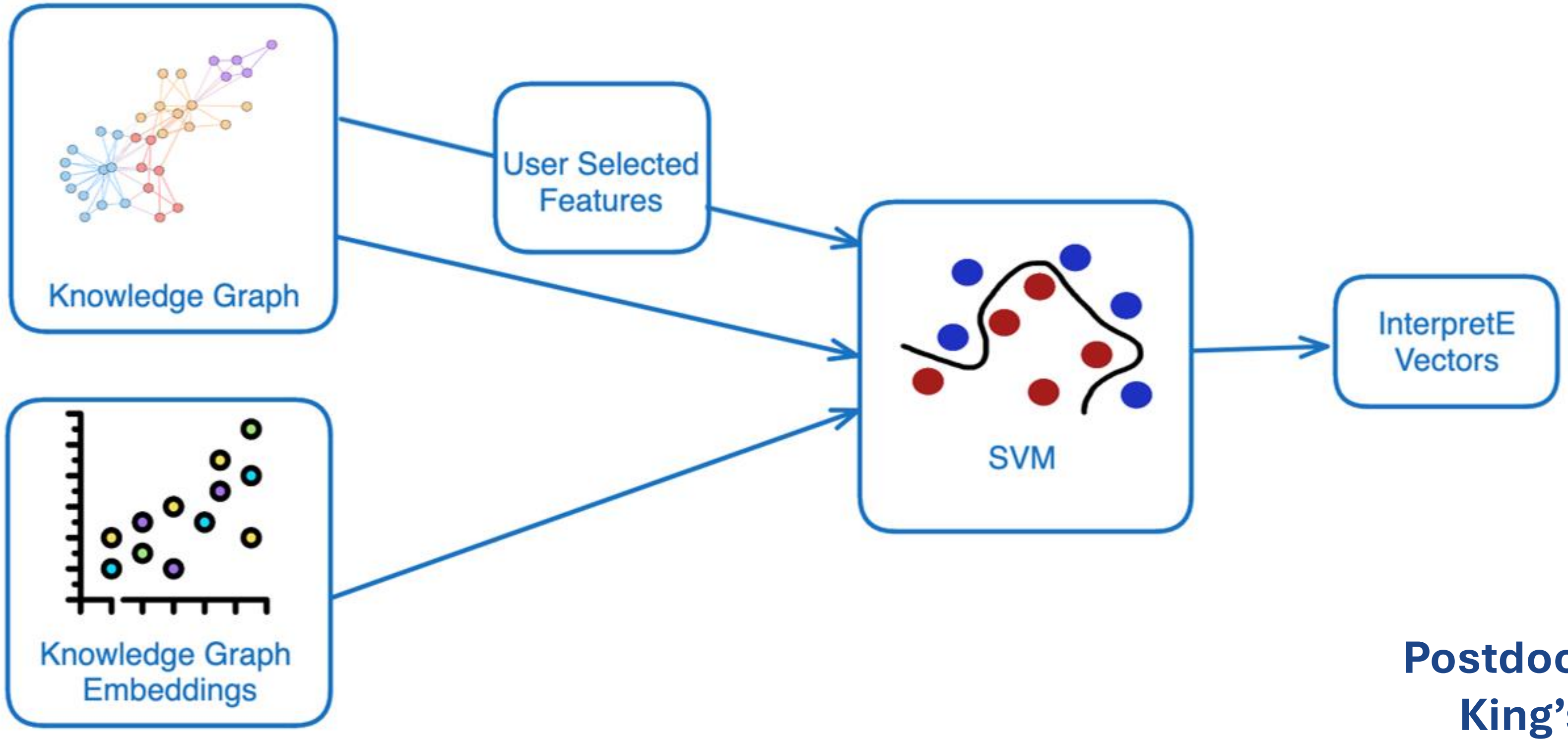
κ scores on the test set and *simtop10* scores on the original and *InterpretE* vectors (mean values) with FB15K-237 for the different KGEMs

Experiments and features		ConvE	TransE	DistMult	Rescal	Complex
person : gender - place_of_birth United States	κ score	.91	.78	.92	.92	.90
	original	.676	.689	.689	.693	.675
	InterpretE	**.909**	**.909**	**.932**	**.99**	**.977**
organizations: locations (USA - UK - Japan - Canada - Germany	κ score	.78	.70	.75	.58	.79
	original	.766	.738	.758	.731	.768
	InterpretE	**.951**	**.947**	**.958**	**.959**	**.96**
film: film_release_region (USA - Sweden - France - Spain - Finland)	κ score	.71	.69	.71	.66	.71
	original	.705	.66	.661	.621	.661
	InterpretE	**.876**	**.866**	**.903**	**.907**	**.892**
film: film genre (drama - comedy - romance - thriller - action)	κ score	.68	.65	.71	.72	.70
	original	.212	.217	.215	.217	.213
	InterpretE	**.732**	**.719**	**.805**	**.78**	**.753**

Conclusions and Future Work

- This work attempts to address the oft overlooked issue of lack of semantic interpretability in latent spaces generated by popular KGE techniques.
- Aim to bridge the gap between entity representations and human-understandable features.
- *InterpretE* approach is capable of deriving interpretable spaces from existing KGEM vectors with human-understable dimensions, based on the features in the underlying KG.
- Future research can further explore the implications of this approach and extend its applicability to broader contexts, more datasets, address scalability of SVM.
- Contribute to the broader goal of AI transparency, ensure that AI-driven systems operate in a manner that is both reliable and ethical.

Thank you for your attention!



Contact

Dr Nitisha Jain

Postdoctoral Research Associate

King's College London (KCL)

nitisha.jain@kcl.ac.uk

Questions and discussion welcome.