Learning Better Representations Using Auxiliary Knowledge

Saed Rezayi

Department of Computer Science, University of Georgia, Athens, GA, USA

The 37th AAAI Conference on Artificial Intelligence – Doctoral Consortium, 2023

February 7, 2023

Introduction

- Representation Learning (RL) is a subfield of machine learning that focuses on learning **meaningful and compact** representations of data.
- RL is a crucial aspect of AI, helping machines to **understand and interpret** complex data inputs.
- With effective RL, AI systems can recognize patterns, make predictions, and **make decisions** based on relevant information.
- **Improving RL** can lead to more accurate and effective AI systems across a variety of applications in CV, NLP, and more.

Enhancing RL

- **Preprocessing of input data**: Cleaning, normalizing, and transforming data.
- **Data augmentation**: Increasing the amount and diversity of training data.
- **Advanced Learning Techniques**: incorporating methods such as transfer learning, multi-task learning, adversarial training, etc.
- **Use of auxiliary knowledge**: Incorporating external knowledge, such as domain-specific information, into the RL system.

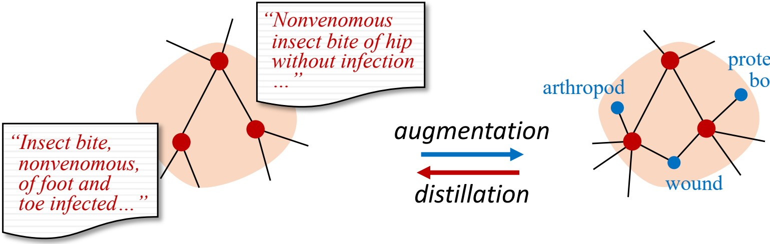
Enhancing RL using Auxiliary Knowledge

1. **Text**: Text data, such as encyclopedias, and other natural language resources, can provide valuable information about the relationships between words, concepts, and entities.
2. **Knowledge Graphs**: Knowledge graphs represent relationships between entities as nodes and edges, providing a rich source of structured knowledge that can be used to improve RL.
3. **Similar Datasets**: Using similar datasets, such as those from related domains or tasks, can provide additional information and insights to improve RL.

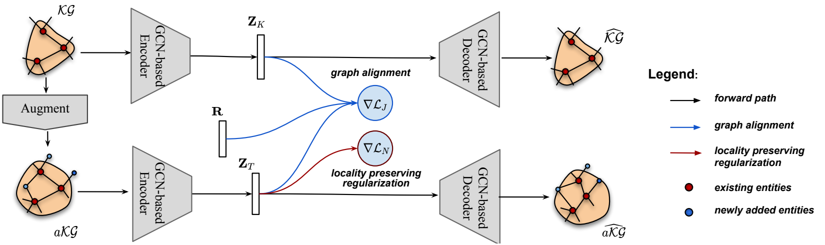
Text as External Source of Knowledge

- **Knowledge Graph Embedding (KGE)**: is a type of RL that maps **entities** and **relationships** in a knowledge graph into a low-dimensional vector space.
- **Knowledge Graphs (KG)**: are a rich source of structured information but are often **sparse**, meaning that there are many missing relationships between entities.
- **External text**: can be used to improve KGE by providing additional information and insights about the relationships between entities in the KG.
- **Research Question**: How can we incorporate text into the learning process?

Augmenting Knowledge Graph



Proposed Method: Edge[1]



[1] Rezayi, Saed, et al. "Edge: Enriching Knowledge Graph Embeddings with External Text." NAACL. 2021.

Enriching KGE

1. Minimize the distance between reconstructed knowledge graph and its original version (MSE):

$$\mathscr{L}_K = \min ||\mathbf{A}_K - \hat{\mathbf{A}}_K||_2$$

2. Align ***KG*** and ***aKG*** through common entities in the embedding space (joint loss):

$$\mathscr{L}_J = ||\mathbf{Z}_K - \mathbf{R}\mathbf{Z}_T||_2$$

3. Force textual nodes related to a target entity to be closer to that entity compared to unrelated ones (negative sampling):

$$\mathscr{L}_N = -\log(\sigma(\mathbf{z}_e^\top \mathbf{z}_t)) - \log(\sigma(-\mathbf{z}_e^\top \mathbf{z}_{t'}))$$

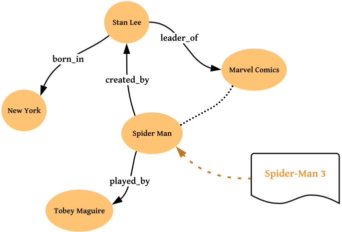
Results

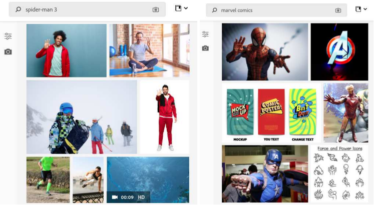
- Link prediction results on SNOMED dataset:

Input	Model	AUC	AP
KG	GAE	0.77	0.84
aKG	GAE	0.86	0.90
KG+aKG	EDGE	**0.91**	**0.94**

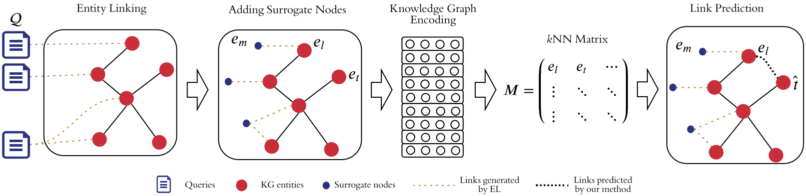
Query Suggestion

- **Search engines** often rely on rich metadata being available for the content items.
- Some search queries might suffer from reduced relevance, i.e., **low precision**.
- **Research Question**: Can we exploit external knowledge to provide entity-oriented reformulation for queries?





Proposed Framework[2]



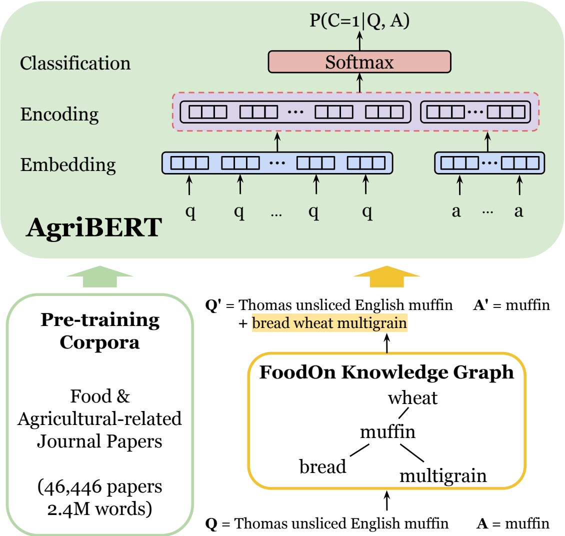
- **INPUT**: A set of queries with unsatisfactory search results.
- **OUTPUT**: A set of KG entities related to input queries.

[2] Rezayi, Saed, et al. "A Framework for Knowledge-Derived Query Suggestions." IEEE BigData, 2021.

Semantic Matching

1. **Problem**: Matching consumer receipt data with USDA nutrition database.
2. We frame the problem as **answer selection** problem.
3. We train a language model using existing literature in agriculture domain including 2.4 million tokens (AgriBERT).
4. **Research Question**: Can we boost the performance of answer selection using an external KG?

Framework[3]



[3] Rezayi, Saed, et al. "Agribert: knowledge-infused agricultural language models for matching food and nutrition." IJCAI, 2022.

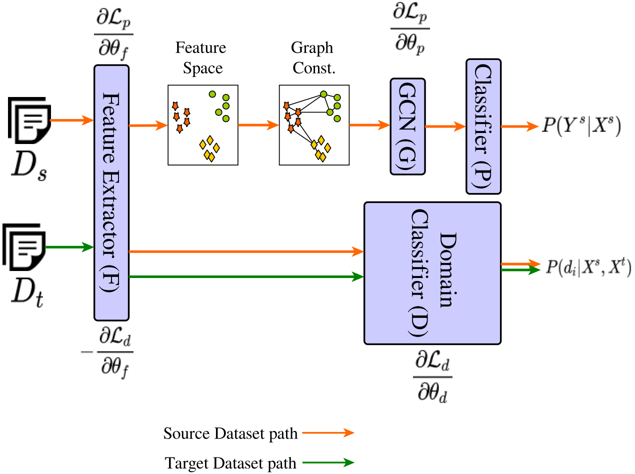
Results

Training Dataset	Model	P@1
-	kNN	14.49
-	BERT	10.88
WikiText-103	BERT+EL (Wikidata)	10.09
WikiText-103	BERT+FoodOn (n=1)	24.83
Agricultural Corpus	BERT	12.71
Agricultural Corpus	BERT+EL (Wikidata)	21.52
Agricultural Corpus	BERT+FoodOn (n=1)	47.89
Agricultural Corpus	BERT+FoodOn (n=3)	49.80
Agricultural Corpus	BERT+FoodOn (n=5)	**49.98**

Cross Domain Clustering

- **Problem**: Short text clustering
- **Challenges**: Unknown data distribution, topic evolution, semantic sparsity.
- **Research Question**: Can we exploit a labeled dataset to model the underlying distribution of a target dataset?

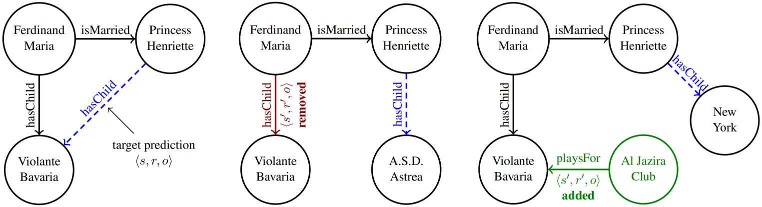
Proposed Model[4]



[4] Rezayi, Saed, et al. "XDC: Adversarial Adaptive Cross Domain short text clustering." SDM, 2023.

Trustworthy KGE

- The embeddings generated by KGE models can be considered a "**black box**" as it is difficult to understand how they are generated and how they contribute in the performance of the model.
- Existing methods try to identify which training facts have been **most influential** to the prediction[5].



(a) Original KG (b) Remove (c) Add

[5] Pezeshkpour, P, et al., "Investigating Robustness of Link Prediction via Adversarial Modifications." NAACL, 2019.

Robust KGE

- To the best of my knowledge, there is no work in the area of robust KGE.
- **Research Question**: can we utilize a form of external knowledge to mitigate the effect of the attack and improve the robustness?
- Using word embedding to enhance the KGE:

$$\hat{\mathbf{e}}_i = \mathbf{e}_i + \mathbf{M}\mathbf{w}_i$$

- Other directions in trustworthy KGE include:
  - Borrowing methods from GNNs.
  - Employing other learning paradigms e.g., Reinforcement Learning.
  - Considering other areas such as Fairness in KGE.

Conclusion

- The incorporation of auxiliary knowledge has proven to be a powerful tool for improving representation learning in AI.
- Different sources of auxiliary knowledge, including text and knowledge graphs, can be used to enhance RL in AI.
- The use of auxiliary knowledge can also be helpful in mitigating the effects of adversarial attacks on knowledge graph embeddings.

For further discussion please contact me at: `saedr@uga.edu`