Global Cross-Lingual Knowledge Graph Entity Alignment with Graph Attention Networks

CS5260 Neural Networks and Deep Learning II - Project Poster

Aaron Low Jin Liang (A0225489R), Bair Ping Hao (A0248309Y),
Chong Hang Kuen Tricia (A0194357U), Ng Gin Wen (A0142949U)

1 Problem Definition

Entity alignment is the task of identifying entities from different knowledge bases that refer to the same real-world object. It has many applications such as information extraction, question answering as well as machine translation.
We would be focusing on entity alignment for the purpose of multilingual machine translation.
We propose using cross-lingual knowledge graph embeddings for cross-lingual entity alignment.
These knowledge graphs would be embedded using a GCN-based encoder network. The performance of our solution would be evaluated against other GCN-based frameworks as benchmark.

2 Dataset

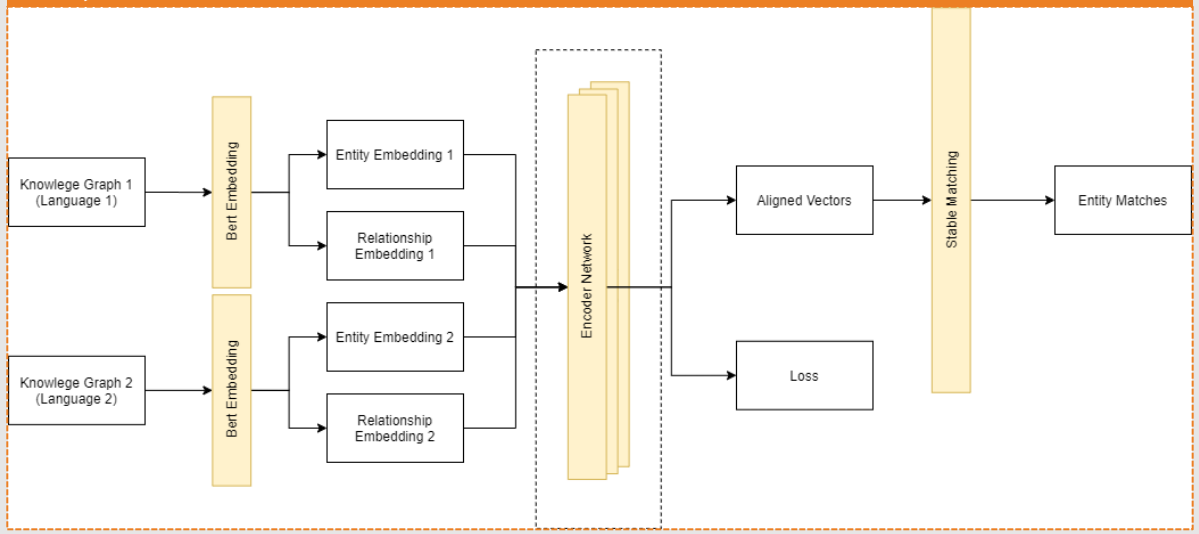
We employed datasets from DBP15K in our project. DBP15K is made up of three different datasets, each corresponding to pairs of knowledge graphs in different languages as shown in the table below. In addition to the knowledge graphs, each dataset also provides 15000 entity pairs between the knowledge graphs.

Dataset	KG1 Nodes	KG1 Edges	KG2 Nodes	KG2 Edges
ZH-EN	19572	70414	19572	95142
FR-EN	19661	105998	19993	115722
JA-EN	19814	77214	19780	93484

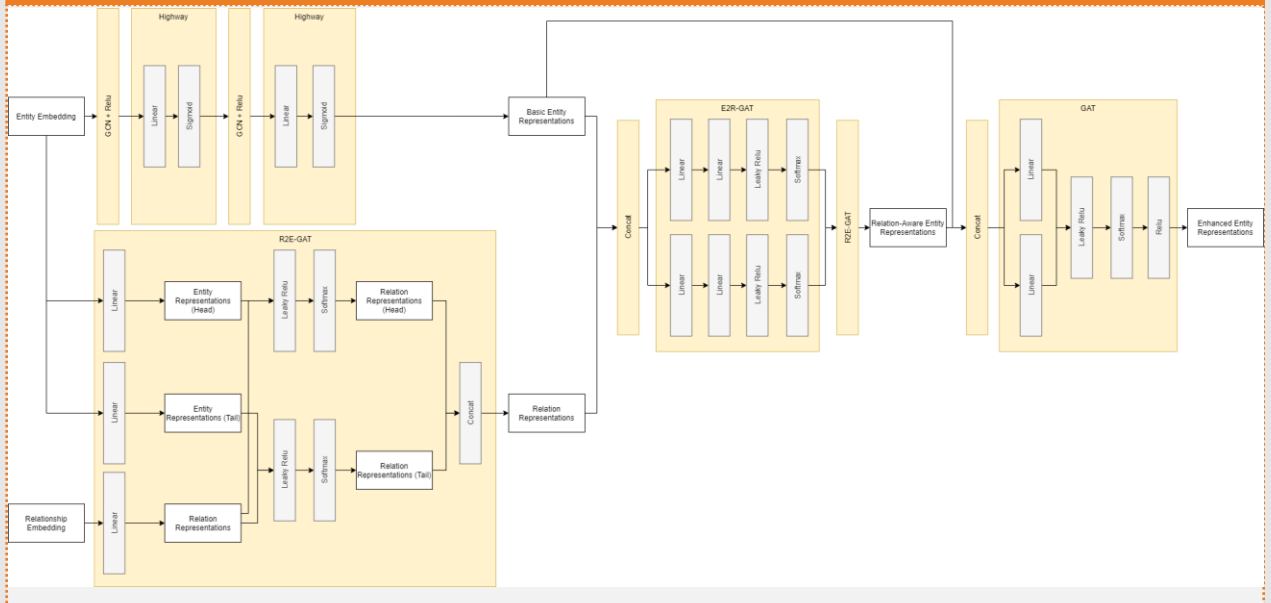
3 BERT Embedding (Pre-Processing)

Previous research uses pre-trained word embedding such as Glove to initialize the entity embeddings. However, this implementation requires translating all entity and relation names into English. Translating non-English entities into English ones and using Glove embedding might lead to the same embedding vectors, undermining the purpose of cross-lingual entity alignment.
In order to build an end-to-end learning pipeline, we utilize a BERT model to embed the names of entities and relations on different knowledge graphs. The pre-trained BERT model is fine-tuned with the downstream task of entity alignment.
More precisely, we first encode the sequences of names and relations into BERT input form and use a pre-trained multilingual BERT model with an additional fully connected layer to retrieve the vector representation of CLS token.
Subsequently, we use a pairwise margin loss to fine-tune the BERT model. Eventually, the fine-tuned BERT model will generate the embedding of multilingual entities and relations in the same vector space.

4 Project Architecture



5 Model Architecture



6 Stable Matching

The output of the GCN-based encoder network is a similarity matrix. Hence, a single entity is similar to multiple entities. Typically, each entity is matched to its most similar counterpart in the other KG. This may result in a one-to-many alignment. However, as our goal is to establish a one-to-one alignment, this assignment task can be framed as a stable matching problem.
The similarity matrix would represent the preferences of each entity, and the goal of the stable matching problem is to assign two entities such that there are no two entities from different KGs that would have preferred another entity other than the assigned entity. The Deferred Acceptance Algorithm was used for this assignment.

7 Experiment Results

All models were trained and evaluated on the JA-EN pair of the DBP15K dataset.

Models	Embed-dings	Hits@1	Hits@10	MRR	Stable Hits
RAGA*	GloVe	83.10%	95.00%	0.875	90.90%
GCN-Align	-	36.3%	68.5%	0.546	-
MTrans E	-	27.9%	57.5%	0.349	-
Modifie d RAGA*	BERT	87.96%	96.86%	0.914	95.46%
TAGCN-Align	GloVe	71.24%	85.93%	0.765	83.88%
TAGCN-Align	BERT	74.56%	87.81&	0.793	87.16%

*RAGA: Relation-Aware Graph Attention

8 Next Steps

We have conducted numerous experiments with different types of embeddings and different model architecture variants.
Based on the experiments that we have conducted, we would pick the model architecture that performs the best and enhance it.

NUS National University of Singapore