Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment

Muhao Chen[1], Yingtao Tian[2], Kai-Wei Chang[1], Steven Skiena[2] and Carlo Zaniolo[1]

[1]University of California Los Angeles; [2]Stony Brook University

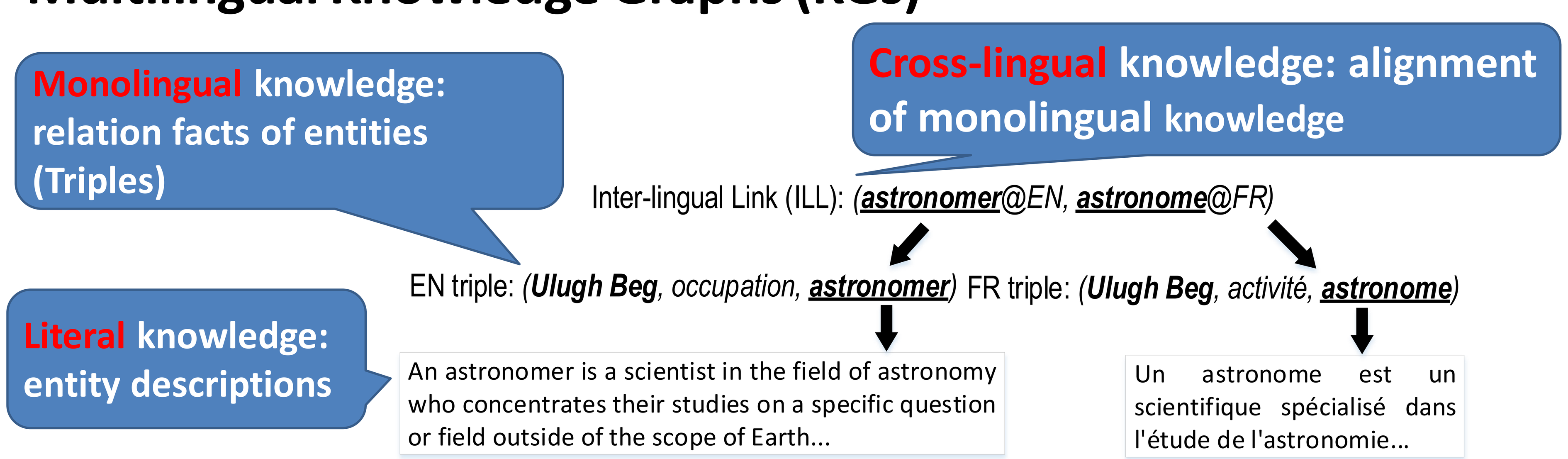
UCLA ENGINEERING Computer Science

Overview

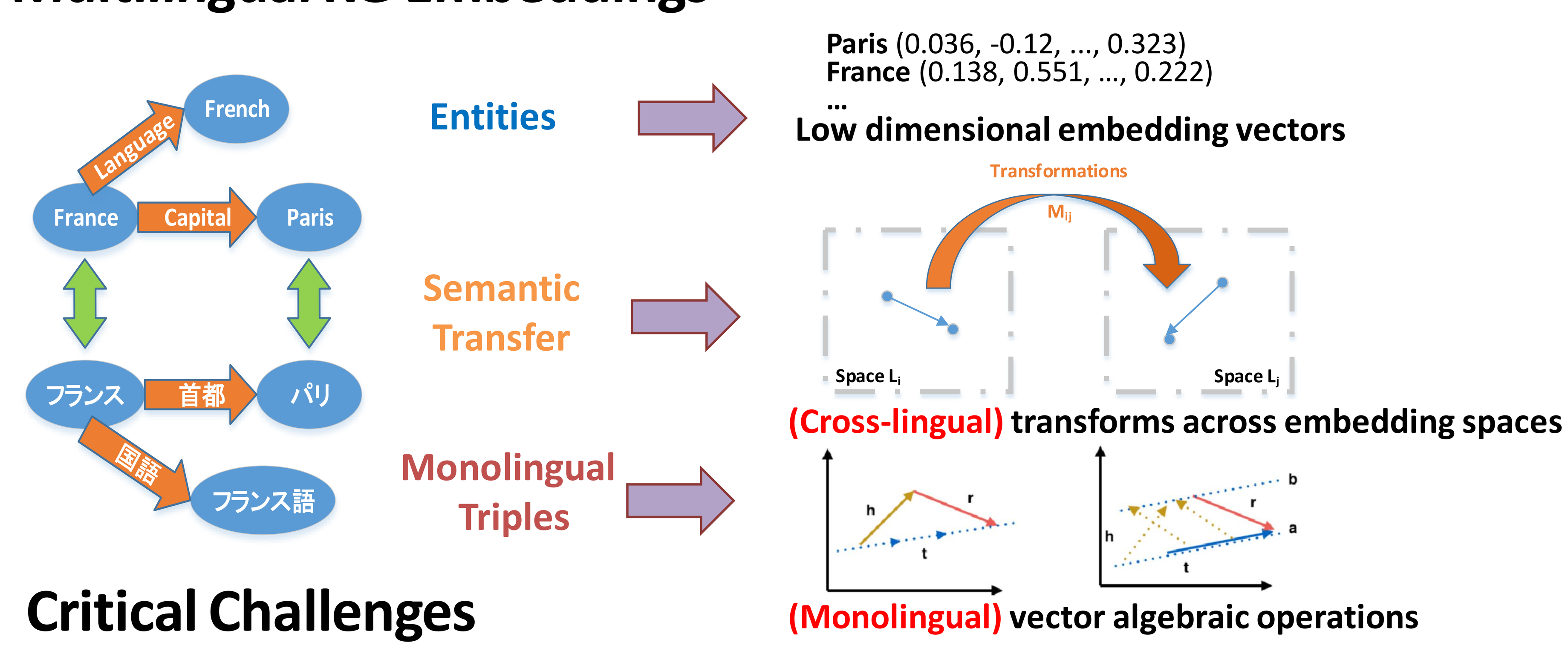
We introduce an embedding-based approach which leverages a weakly aligned multilingual KG for semi-supervised cross-lingual learning using entity descriptions. Our approach performs co-training of two embedding models, i.e. a multilingual KG embedding model and a multilingual literal description embedding model. The models are trained on a large Wikipedia-based trilingual dataset where most entity alignment is unknown to training. Experimental results show that the performance of the proposed approach on the entity alignment task improves at each iteration of co-training, and eventually reaches a stage at which it significantly surpasses previous approaches. We also show that our approach has promising abilities for zero-shot entity alignment, and cross-lingual KG completion.

Preliminaries

Multilingual Knowledge Graphs (KGs)



Multilingual KG Embeddings



Critical Challenges

- **Weak alignment**: existing approaches rely on seed alignment of graph structures to learn cross-lingual semantic transfer, which is insufficiently populated in many large knowledge bases.
- **Zero-shot scenarios**: existing approaches represent cross-lingual entities solely on KG structures, which cannot represent entities that do not connect to the structure

Proposed Approach: KDCoE

- Embedding KG and Entity Descriptions for semi-supervised cross-lingual learning
- Iteratively co-training two model components:
  1. Multilingual KG embedding model (KGEM)
  2. Multilingual entity description embedding model (DEM)

Multilingual KG Embedding Model

- MTransE-LT
  1. KG Structure Encoder

$$S_K = \sum_{L\in\{L_i,L_j\}} \sum_{(h,r,t)\in G_L \wedge (\hat{h},r,\hat{t})\notin G_L} \left[f_r(h,t) - f_r(\hat{h},\hat{t}) + \gamma\right]_+$$

$$f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2$$

  2. Jointly-trained alignment model

$$S_A = \sum_{(e,e')\in I(L_i,L_j)} \left\|\mathbf{M}_{ij}\mathbf{e} - \mathbf{e}'\right\|_2$$

- Learning Objective

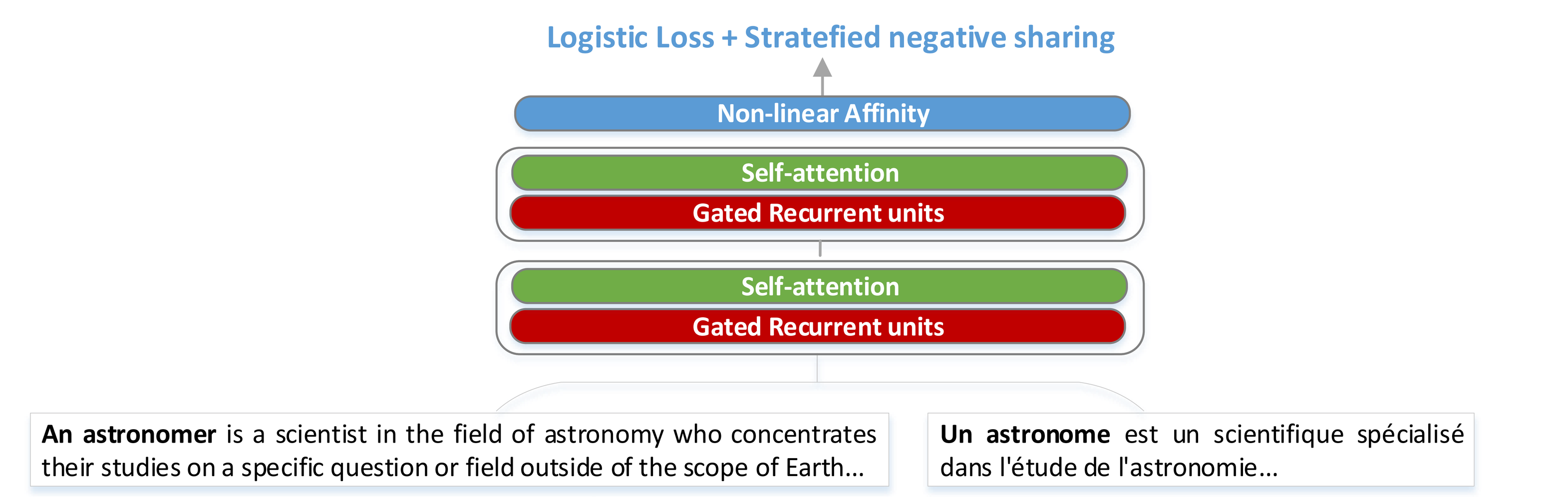
$$S_{KG} = S_K + \alpha S_A$$

Data	#En	#Fr	#De	ILL Lang	#Train	#Valid	#Test	#Zero-shot
Triples	569,393	258,337	224,647	En-Fr	13,050	2,000	39,155	5,000
Desc.	67,314	45,842	43,559	En-De	12,505	2,000	41,018	5,632

Table 1: Statistics of the Wk3l60k dataset.

Entity Description Embedding Model

- Siamese GRU Encoder with Self-attention



- Logistic loss with stratified negative sharing (shared negative samples in one batch)

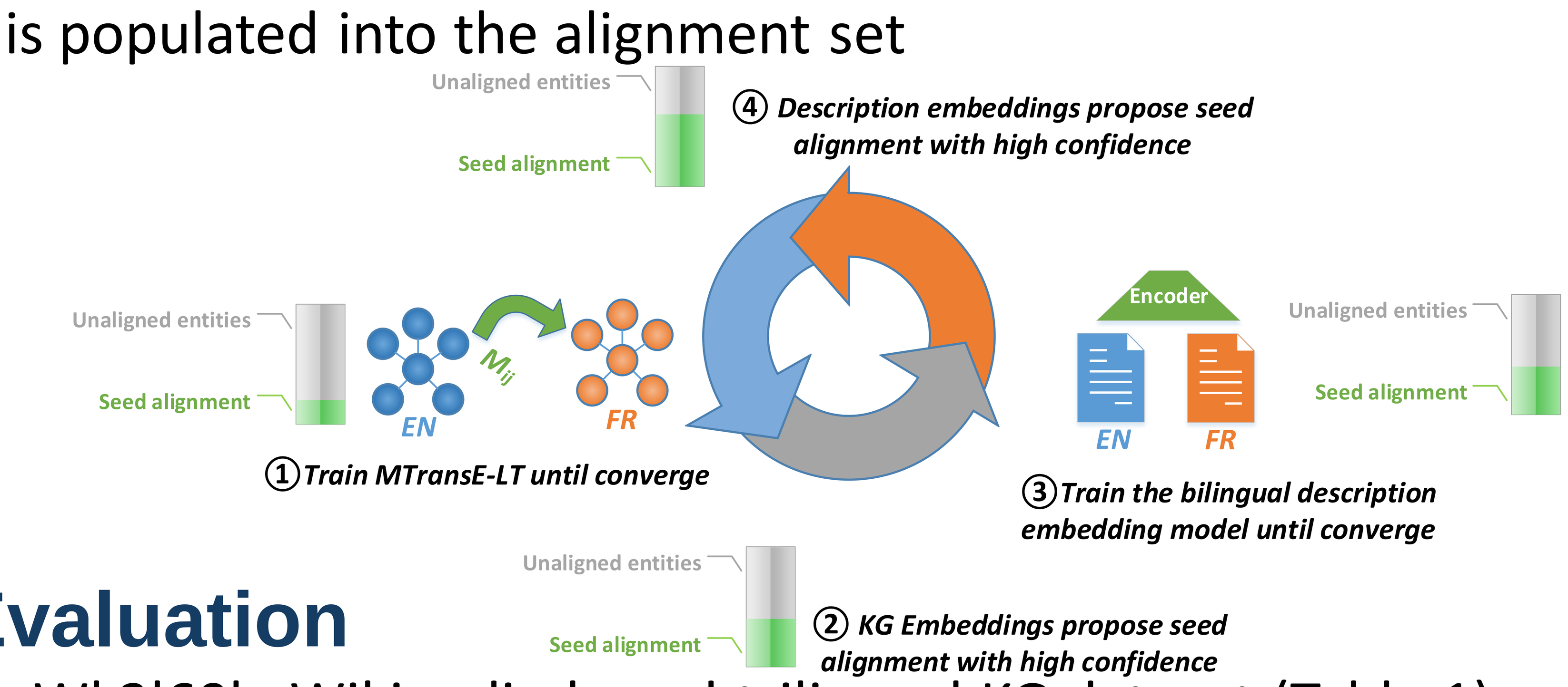
$$S_D = \sum_{(e,e')\in I(L_1,L_2)} -LL_1 - LL_2$$

$$LL_1 = \log\sigma(\mathbf{d}_e^\top \mathbf{d}_{e'}) + \sum_{k=1}^{|B_d|} \mathbb{E}_{e_k\sim U(e_k\in E_{L_i})}\left[\log\sigma(-\mathbf{d}_{e_k}^\top \mathbf{d}_{e'})\right]$$

$$LL_1 = \log\sigma(\mathbf{d}_e^\top \mathbf{d}_{e'}) + \sum_{k=1}^{|B_d|} \mathbb{E}_{e_k\sim U(e_k\in E_{L_j})}\left[\log\sigma(-\mathbf{d}_e^\top \mathbf{d}_{e_k})\right]$$

Iterative co-training

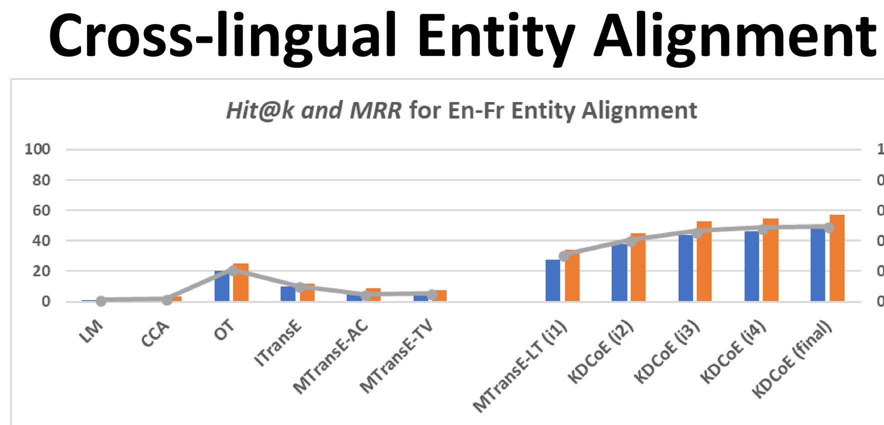
- Iteratively co-training two model components based on the growing alignment set
- Seed alignment with high confidence (low embedding distance) is populated into the alignment set

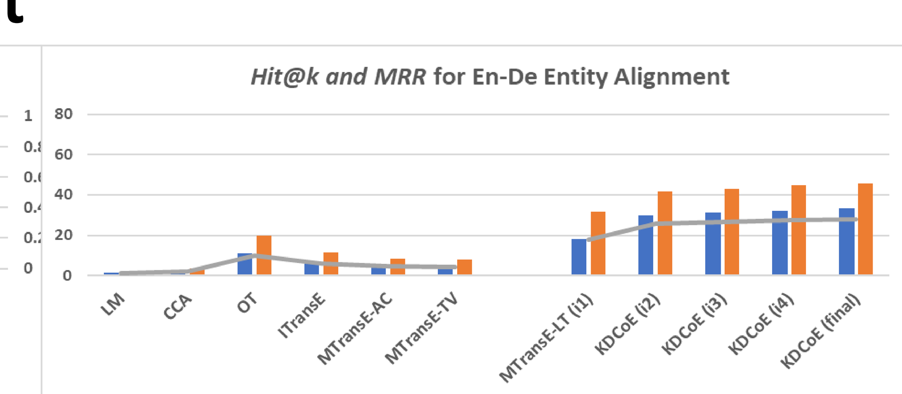


Evaluation

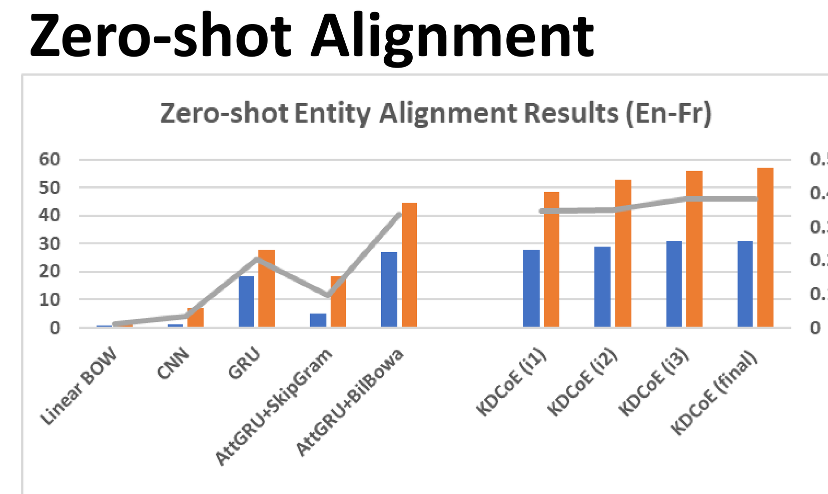
- Wk3l60k: Wikipedia-based trilingual KG dataset (Table 1)

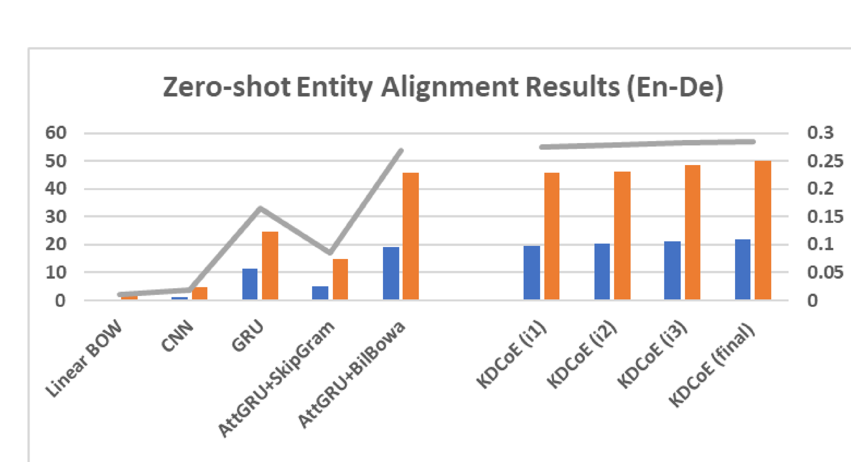
Cross-lingual Entity Alignment





Zero-shot Alignment





Cross-lingual KG Completion