Taxonomy Construction of Unseen Domains via Graph-based Cross-Domain Knowledge Transfer

Chao Shang, Sarthak Dash, Md Faisal Mahbub Chowdhury,
Nandana Mihindukulasooriya, Alfio Gliozzo

ACL2020

Problem definition

Taxonomy: classify things into hierarchical structures e.g. graph/tree

Examples: wordnet

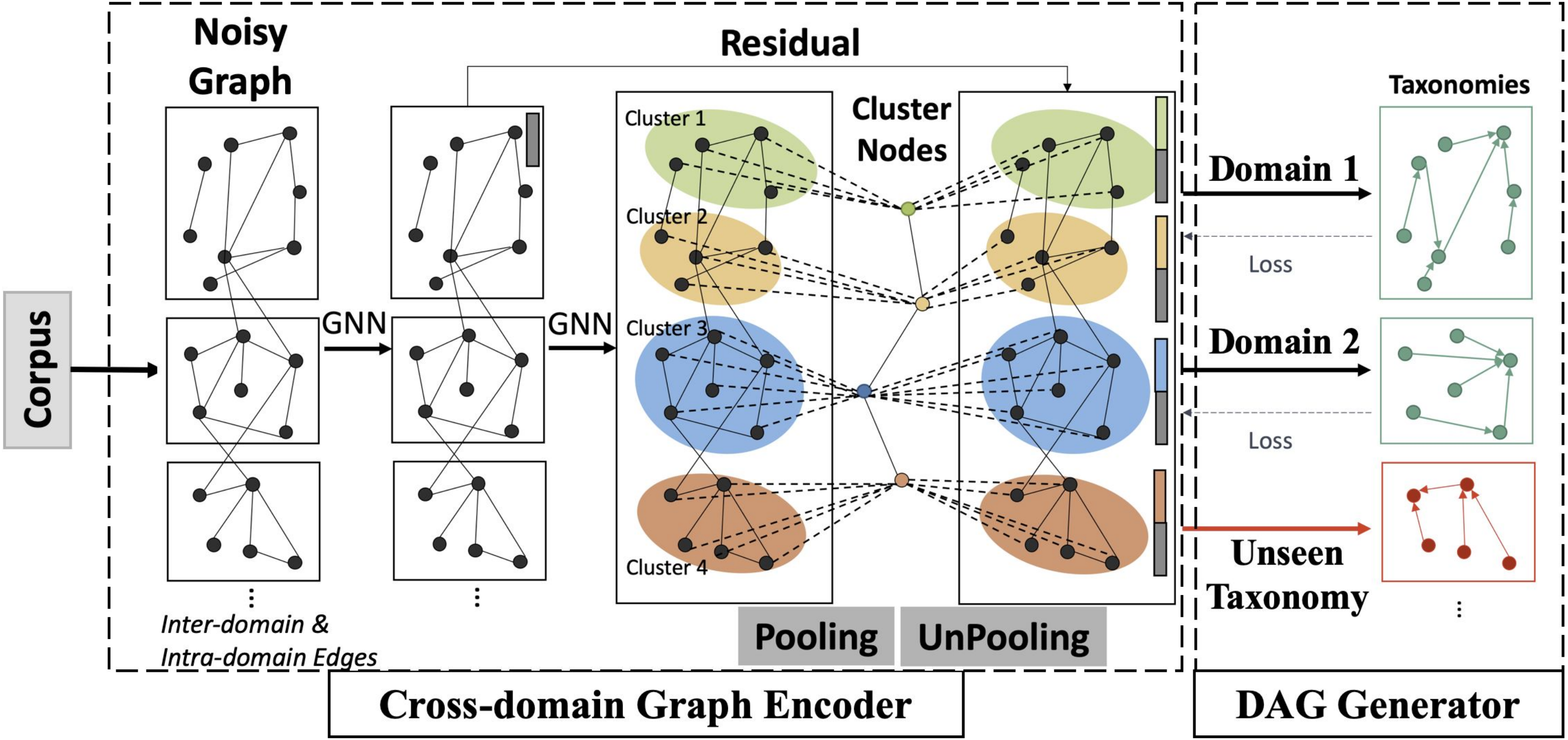
The problem: Given a list of domain-specific terms from a target unseen domain as input, how to construct a taxonomy for that target unseen domain

Or: given a list of terminologies in an unseen domain, how to organize them into a taxonomy

Problem setting

- Train set:
  - A large corpus
  - A set of golden taxonomies from some known domains
- Testing set
  - An unseen corpus
  - A set of terminologies of target unknown domain
- Output
  - A taxonomy of the target unknown domain

Framework



Corpus
Noisy Graph
GNN
GNN
Residual
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster Nodes
Inter-domain &
Intra-domain Edges
Pooling
UnPooling
Taxonomies
Domain 1
Loss
Domain 2
Loss
Unseen Taxonomy
Cross-domain Graph Encoder
DAG Generator

Build cross-domain noisy graph

- Extract candidates: “is-a” pairs from a large collection of input using substring matching and pattern-based approaches
  - This graph is very noisy
  - ...animals other than dogs such as cats...
  - -> (cat, is-a, dog)
- Subgraph extraction $G_{input} = (V_{input}, E_{input})$
  - V_input is a set of interested terms
  - E_input contains (v_i, v_j) if (v_i, v_j) appear in the noisy graph

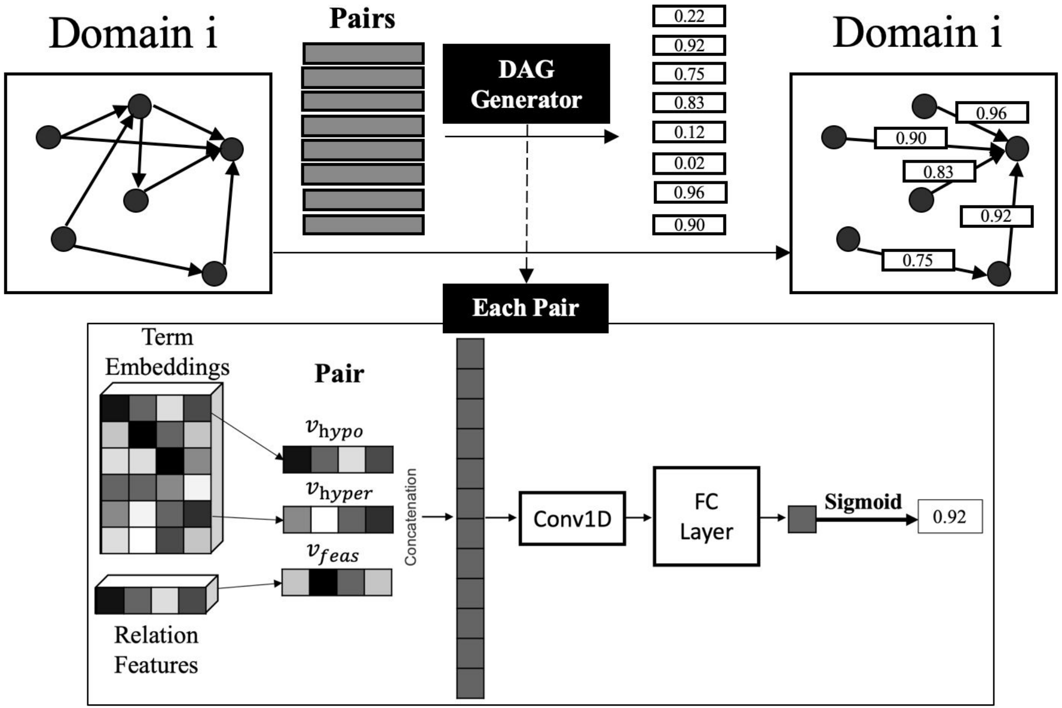
Cross domain graph encoder

- Neighbor aggregation $H^{l+1} = GNN_l(A, H^l) = \sigma(\tilde{A}H^l\Theta^l)$
- Semantic Clustering Aggregation
  - Generate soft assignment $S^l = softmax(GNN_{l,cluster}(A, H^l))$
  -
  - Generate cluster embedding Hc $H_c^l = (S^l)^T H^l \in \mathbb{R}^{n_c \times d_l}$
  -
  - Generate cluster graph $A_c = (S^l)^T A S^l \in \mathbb{R}^{n_c \times n_c}$
  -
  - Forward through cluster graph $H_c^{l+1} = GNN_l(A_c, H_c^l) \in \mathbb{R}^{n_c \times d_{l+1}}$
  -
  - Unpooling the cluster embedding to restore the original graph $\tilde{H}^{l+1} = S^l H_c^{l+1} \in \mathbb{R}^{n \times d_{l+1}}$
- Combine representation

$$H^{l+1} = concate(\tilde{H}^{l+1}, H^l)$$

Link prediction

$$v_{pair} = concate(v_{hypo}, v_{hyper}, v_{feas})$$
$$p_{(hypo,hyper)} = sigmoid(V_C^T W)$$



Results

	Science (Combined)			Science (Eurovoc)			Science (WordNet)			Science (Average)			Environment (Eurovoc)		
Model	P_e	R_e	F_e	P_e	R_e	F_e	P_e	R_e	F_e	P_e	R_e	F_e	P_e	R_e	F_e
Baseline	0.63	0.29	0.39	0.62	0.21	0.31	0.69	0.27	0.38	0.65	0.26	0.36	0.50	0.21	0.30
JUNLP	0.14	0.31	0.19	0.13	**0.36**	0.19	0.21	0.31	0.25	0.16	0.33	0.21	0.13	0.23	0.17
USAAR	0.38	0.26	0.31	0.63	0.15	0.25	**0.82**	0.19	0.31	0.61	0.20	0.29	0.81	0.15	0.25
TAXI	0.39	**0.35**	0.37	0.30	0.33	0.31	0.37	**0.38**	0.38	0.35	0.35	0.35	0.34	0.27	0.30
TaxoRL^A	–	–	–	–	–	–	–	–	–	0.57	0.33	0.42	0.38	0.24	0.29
TaxoRL^B	–	–	–	–	–	–	–	–	–	0.38	**0.38**	0.38	0.32	0.32	0.32
Graph2Taxo^1	**0.91**	0.31	0.46	0.78	0.26	0.39	**0.82**	0.32	**0.46**	**0.84**	0.30	0.44	**0.89**	0.24	0.37
Graph2Taxo^2	0.90	0.33	**0.48**	**0.79**	0.33	**0.46**	0.77	0.32	**0.46**	0.82	0.33	**0.47**	0.67	**0.28**	**0.39**