HAETAE: In-domain Table Pretraining with Header Anchoring

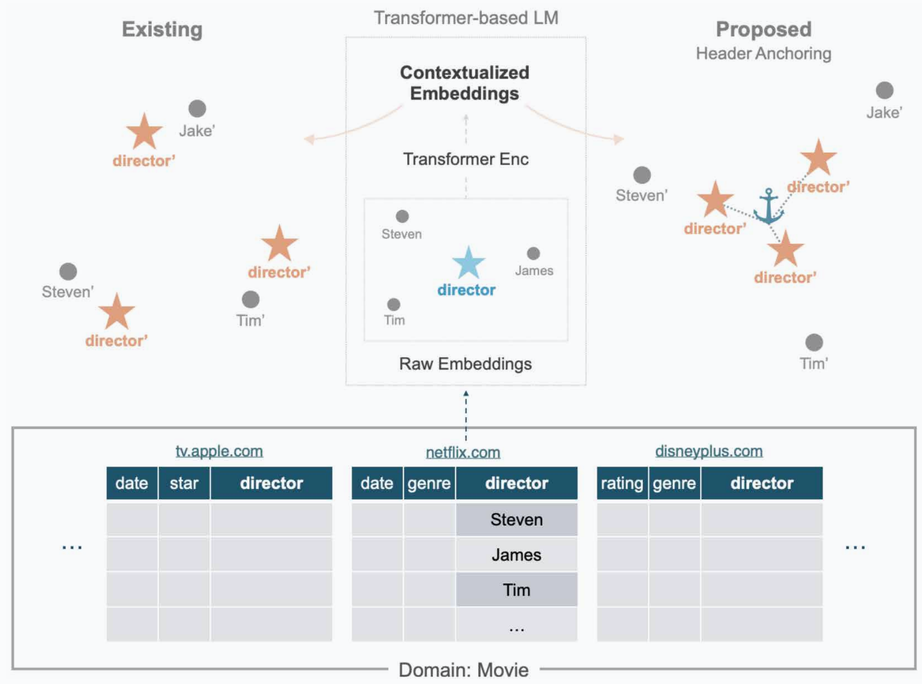
Woojun Jung, Susik Yoon

Korea University, Seoul, South Korea

Contact woojoon@korea.ac.kr | **Github** https://github.com/woojoonjung/HAETAE

Introduction

- **Tabular data is central** to tasks in information retreival such as QA, reasoning, classification
- Transformer-based table embedding models (e.g. TaBERT, TaPas) face key limitations:
  - They **flatten tables into sequences**, losing structural relationships
  - They rely on **fully contextualized representations**, making them **inconsistent across tables with similar headers**
  - They fail to stably encode header semantics, reducing generalization to new entities or tables
- Real-world tables often has **semi-homogeneous headers** across domain-specific tables
- Without stable header representation, models cannot generalize across in-domain tables
- Our goal is to **preserve universal semantics of headers across entities and tables**
  enable stable, transferable representations for structured data



Methodology

- **Header-aware serialization** converts tabular data into a dictionary format to preserve header-value pairs explicitly and handle nested structures
- This input format explicitly enables schema preservation
- **Universal header encoder** is dedicated to learning stable header embeddings
- This decouples header representations from surrounding entity context
- **Universal-contextual alignment** is utilized to align universal header embeddings with BERT's contextual embeddings using:
  - *Interpolation*: Blends universal and contextual header embeddings and calculate MLM loss

$$\hat{\mathbf{E}}^{t_i} = \begin{cases} \mathbf{E}^{t_i}_{\text{inter}}, \text{ if } t_i \text{ is a header token} \\ \mathbf{E}^{t_i}_{\text{cont}}, \text{ otherwise.} \end{cases}$$

$$\mathbf{E}^{t_i}_{\text{inter}} = \alpha \cdot \mathbf{E}^{t_i}_{\text{cont}} + (1-\alpha) \cdot \mathbf{E}^{t_i}_{\text{univ}}$$

(α is a learnable parameter)

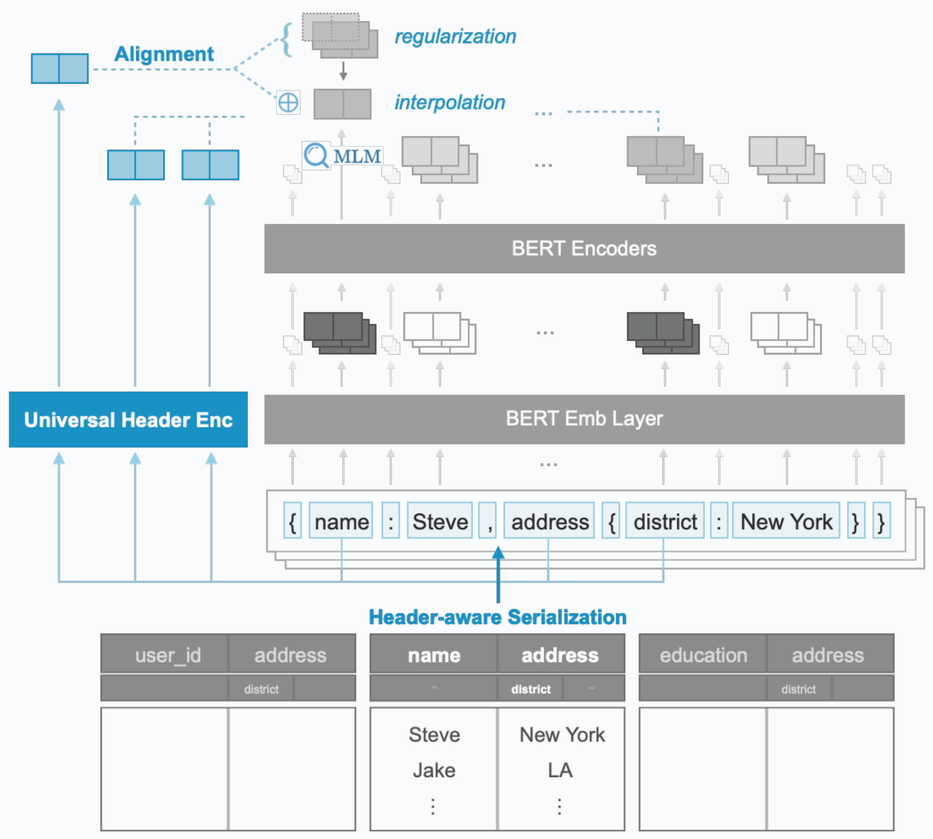
$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{M}|} \sum_{t_i \in \mathcal{M}} \log P(t_i | \mathcal{T}_i),$$

  - *Regularization*: Encourages batch-wise consistency of header embeddings by minimizing distance between universal embeddings and contextual embeddings across entity samples

$$\mathbf{C}^{t_i}_{\text{cont}} = \frac{1}{|\mathcal{O}_{t_i}|} \sum_{t_{i,j} \in \mathcal{O}_{t_i}} \mathbf{E}^{t_{i,j}}_{\text{cont}}$$

$$\mathcal{L}_{\text{header}} = \frac{1}{|\mathcal{K}|} \sum_{t_i \in \mathcal{K}} \left| \mathbf{C}^{t_i}_{\text{cont}} - \mathbf{E}^{t_i}_{\text{univ}} \right|_2^2,$$

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{header}}.$$



Experiments

Masked Header/Value Prediction

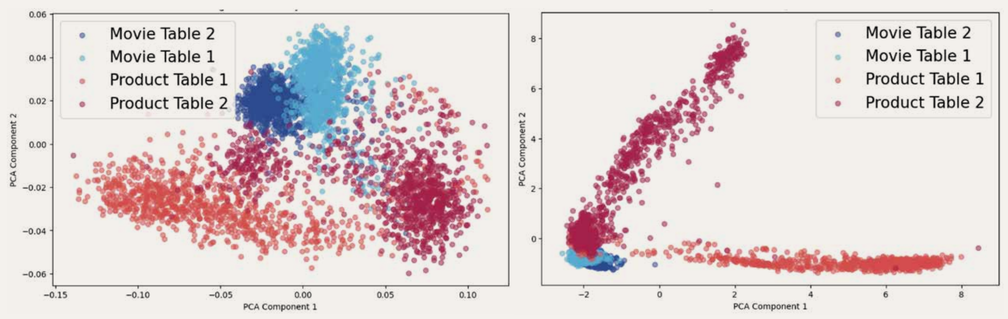
Quantitative evaluation of semantic generalization on two in-domain datasets

Model	Product		Movie	
	Header	Value	Header	Value
$\text{BERT}_{\text{base}}$	0.4382	0.5701	0.5651	0.5929
TaPas	0.2326	0.3807	0.2625	0.3932
TaBERT	0.4479	0.5661	0.5596	0.5957
$\text{BERT}_{\text{dspt}}$	0.9009	0.6823	0.8744	0.6996
$\text{HAETAE}_{\text{dspt}}$	**0.9923**	**0.7081**	**0.9942**	**0.7271**
-w/o Interpolation	0.5642	0.7056	0.5830	0.7192
-w/o Regularization	0.9926	0.7068	0.9922	0.7188

- HAETAE outperforms both general (BERT) and table-specific (TaBERT, TaPas) baselines
- Ablation shows that removing interpolation significantly harms header prediction
- Both interpolation and regularization contribute to value prediction performance

Visulization of Entity Embeddings

Qualitative analysis using PCA projections of mean-pooled entity representations from 2 tables from each domain (product, movie)



- HAETAE embeddings (left) form:
  - Compact, well-separated clusters within and across domains
  - More stable alignment between similar schemas
- BERT embeddings (right) form:
  - Dispersed clusters, poor separation of domains
  - High variance across tables with similar headers

Conclusion

- HAETAE improves both token-level prediction performance and entity-level alignment
- HAETAE shows strong generalization to unseen entities and schemas
- HAETAE captures structural consistency and semantic coherence across tables better than baselines
- Results demonstrate effectiveness of header anchoring and alignment mechanisms