Budget-Aware Entity Matching Across Domains

Nick Pulsone, Prof. Roee Shraga, Greg Goren
Email: nbpulsone@wpi.edu

Introduction & Motivation

TITLE	BRAND	PRICE
Kingston ValueRAM 4 GB DDR4-2400	Kingston	€ 21.95
TP LINK 20000mAh Power Bank	None	€37.29
Cap New Era NY Yankees MLB 9Fifty Cap	None	3.19E1

(a) (b) (c)

TITLE	BRAND	PRICE
Kingston SO-DIMM DDR4 4Gb2400MHz (KVR24S17S6/4)	Kingston	None
Cap New Era NY Yankees 39Thirty Cap	None	2.59E1
Power Bank TP-LINK TL-PB20000	TP-Link	120.90 zł

- **Entity Matching (EM):** Determines whether two records refer to the same real-world entity; commonly used in data integration.

SELECTED TRAINING SET

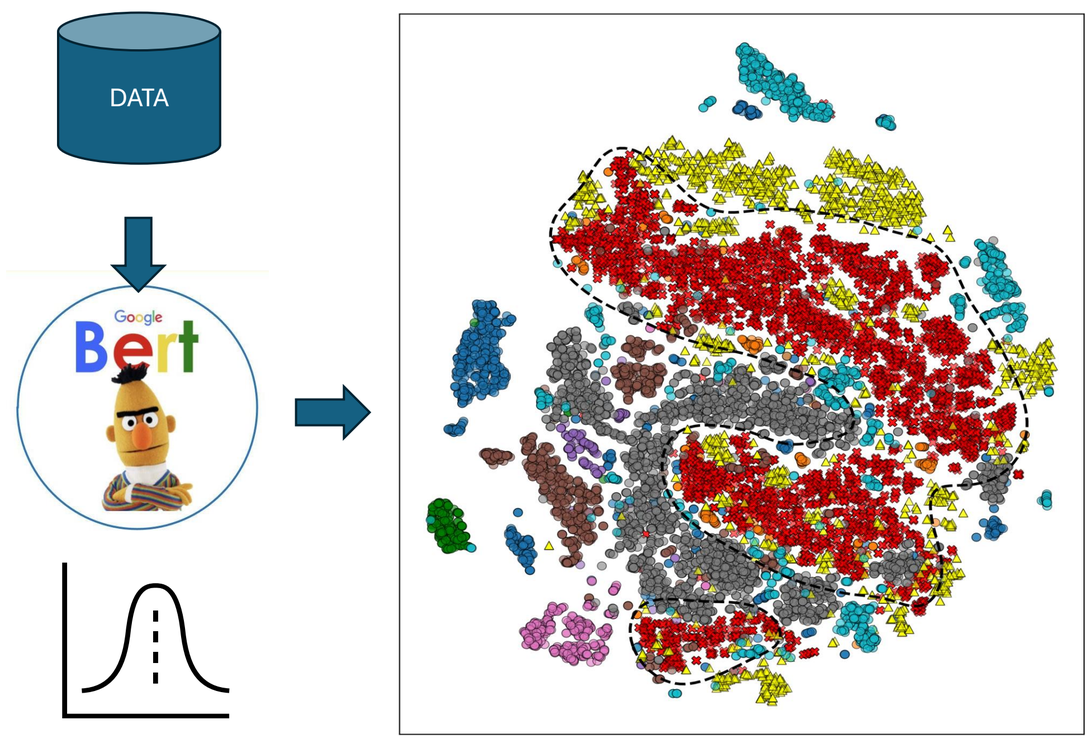
(a) COMPUTERS

TITLE	BRAND	PRICE	TITLE	BRAND	PRICE
Kingston ValueRAM 4 GB DDR4-2400	Kingston	$321.01	Kingston SO-DIMM DDR4 4Gb2400MHz (KVR24S17S6/4)	Kingston	$74.17

	TITLE	BRAND	PRICE	TITLE	BRAND	PRICE
(b) ELECTRONICS	TP LINK 20000mAh Power Bank	None	€37.29	Power Bank TP-LINK TL-PB20000	None	120.90 zł
(c) CLOTHING	Cap New Era NY Yankees MLB 9Fifty Cap	None	3.19E1	Cap New Era NY Yankees 39Thirty Cap	None	2.59E1

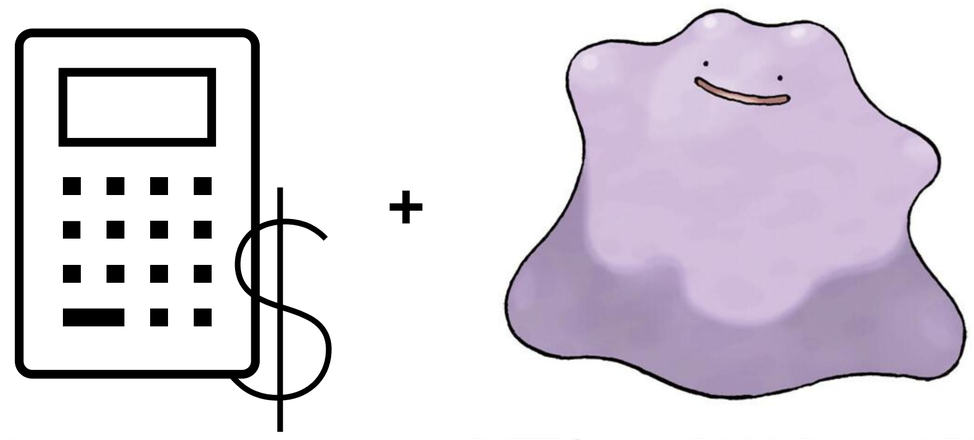
- **Domain-Aware EM:** Trains domain-specific models (e.g., product categories) using both **in-domain** and selected **out-of-domain** data.
- **Key challenge:** Selecting the most *effective* training samples for a target domain.

Distribution-Aware Sample Selection



- Embedding spaces form **clustered regions** that reflect semantic similarity.
- **BEACON** exploits this geometry for informed sample selection. It uses an ensemble of two selection methods:
  - **K-Center Greedy (KCG)**
  - **Train-Validation Distribution Fitting (TVDF)**

The BEACON Model



Budget-Aware + **DITTO [1]: PLM-Based EM**

- **BEACON:** Distribution-aware, budget-aware framework for low-resource EM.
- Guides out-of-domain sample selection using **embedding representations** of record pairs
- Operates under a **fixed annotation budget** for model fine-tuning.

Experiments & Results

- **Dataset:** WDC Multi-Dimensional EM Benchmark [4]
- **Budgets:** 1k-10k training samples
- **Baselines compared:**
  - **SPEC:** Finetuning with domain-specific data only
  - **GEN:** Finetuning with random samples
  - **MFSN [3]:** The SOTA cross-domain EM method
  - **LLAMA [5]:** A zero-shot LLLM baseline for EM
  - **JELLYFISH [6]:** A fine-tuned LLM for EM

Method	5.0k	10.0k	Mean	SD
SPEC	0.655	0.688	0.660	0.034
GEN	0.660	0.709	0.656	0.057
BEACON (ours)	**0.758**	**0.769**	**0.752**	0.025
MFSN [2]	0.658	0.631	0.637	0.020
LLAMA [5]	0.659	0.659	0.659	0.000
JELLYFISH [6]	0.692	0.692	0.692	0.000

Related Work		Domain-Aware	Budget-Aware
1	Deep Entity Matching with Pre-Trained Language Models (DITTO)	✗	✗
2	Deep Entity Matching with Adversarial Active Learning (DEAM)	✗	✓
3	Matching Feature Separation Network (MFSN)	✓	✗
BEACON (ours)		✓	✓

References

[1] Yuliang Li et al. 2020. Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment 14, 1 (2020), 50–60.
[2] Jiacheng Huang et al. 2023. Deep entity matching with adversarial active learning. The VLDB Journal 32, 1 (2023), 229–255.
[3] Chenchen Sun, et al. 2024. Matching feature separation network for domain adaptation in entity matching. In Proceedings of the ACM Web Conference 2024. 1975–1985.
[4] Ralph Peeters et al. WDC Products: A Multi-Dimensional Entity Matching Benchmark. arXiv:2301.09521
[5] Abhimanyu Dubey et al. 2024. The llama 3 herd of models. arXiv e-prints (2024), arXiv–2407.
[6] Haochen Zhang et al. 2023. Jellyfish: A large language model for data preprocessing. arXiv preprint arXiv:2312.01678 (2023)