Product Entity Matching via Tabular Data

Ali Naeim abadi, Mir Tafseer Nayeem, and Davood Rafiei
University of Alberta

Product Entity Matching (PEM)

- **Product Entity Matching (PEM)** is a subfield of record linkage that focuses on linking records that refer to the same product.

Applications

✓ Price Comparison
✓ Comprehensive Catalog
✓ Efficient Data Management
✓ Increased Sales

Challenges

- Multiple features of a product may **be packed into a product title**.
- Some product titles are **highly similar** but are labeled as **non-matching pairs**.
- Pieces of information may be in **different places** for different products.
- Existing datasets, a **fixed number of attributes** are given for all samples.

Amazon Product Title	Google Product Title
"mcafee total protection 2007 3 users"	"mtp07emb3rua mcafee total protection 2007 complete package 3 users cd mini-box"
"britannica deluxe"	"britannica deluxe 2008"
"nero 7 ultra edition enhanced"	"70009 nero ultra edition enhanced v.7 complete package 1 user cd win"

Table 1: A few hard negative examples [20]. Despite their highly similar titles, product pairs are not the same.

Contributions

TAble & Text for Entity Matching (TATEM)
Attribute Ranking Module (ARM)

- Enrich **popular and challenging** benchmarks with complementary product tables.
- Propose a **new serialization technique** to encode semi-structured tables.
- TATEM employs both tabular and textual information **reaching a new SOTA.**
- Design ARM to **select important product-specific attributes** and to make the model data-efficient

Dataset

- We enriched popular PEM datasets.
- We added product-specific tables
  - **Varying numbers** of attributes
  - Many **distinct schemas.**

	Amazon-Google [20]	Walmart-Amazon [20]
#Train samples (N./P.)	6175/699	5568/579
#Test samples (N./P.)	2059/234	1856/193
#Attrs (*fixed*)	3	5
	Amazon-Google-Tab (*ours*)	**Walmart-Amazon-Tab** (*ours*)
#Tables (Amazon)	909	16264
Table coverage	66%	73%
#Unique attrs	84	695
Avg. #attrs	10.2	19.97
Max #attrs	28	81

Table 2: Statistics of the datasets.

TATEM Model

TATEM Serialization

TATEM employs **a serialization technique** for semi-structured, product specific data.

$$e = (title,\ val_{title}), (manufac,\ val_{manufac}), (price,\ val_{price}), \{(attr_i,\ val_i)\}_{1\le i\le k}.$$

$$serialize(e) ::= val_{title}\ [\text{ATTR}]\ val_{manufac}\ [\text{ATTR}]\ val_{price}\ [\text{ATTR}]\ (attr_i,\ val_i)\ldots[\text{ATTR}]\ (attr_k,\ val_k)).$$

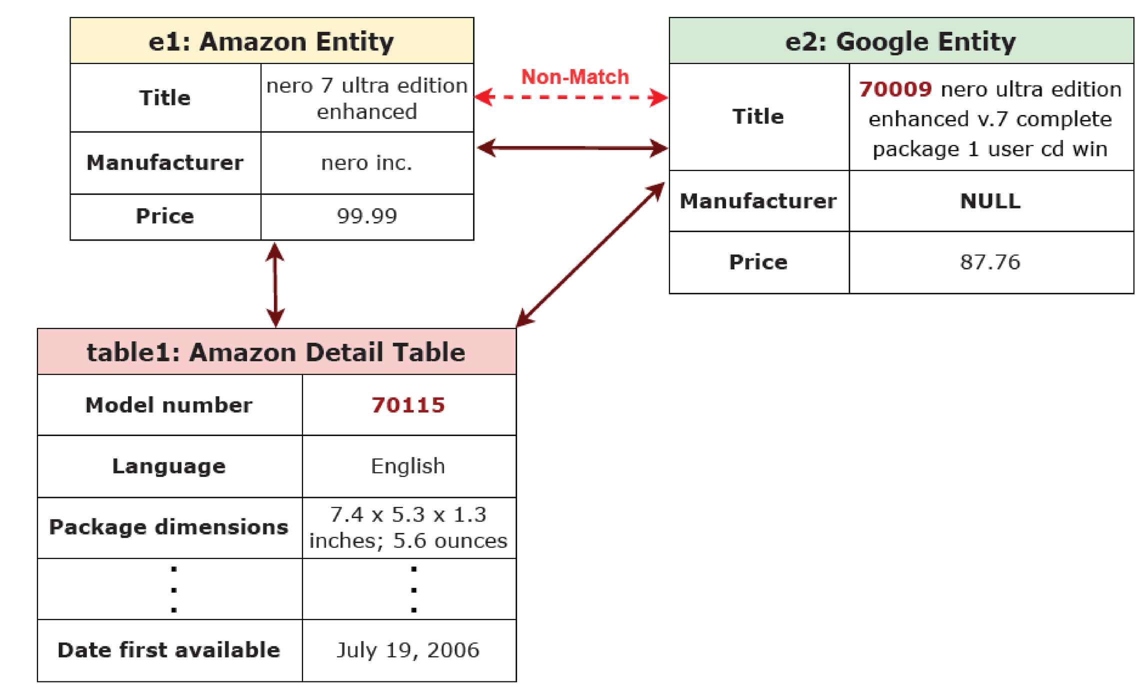

Figure 1: A hard negative example disambiguated using an Amazon product detail table, showing that relying on the information given in titles alone is hard to vote against a match because of the large number of overlapping tokens. Our model TATEM disambiguates this by establishing a relationship between $e2$ and $table1$ (*if exists*). Here, the `Model Number` field helps TATEM to reach a `Non-Match` decision.

Attribute Ranking Module (ARM)

Generate **the top n attribute-value pairs** for a given product entity (e.g., from Amazon) in response to a pair of entities (e.g., Amazon-Google) for EM.

- **Three Benefits**
  - An **effective solution** for transformer-based PLMs on **length limitation**.
  - TATEM equipped with ARM improves the overall efficiency and **effectiveness of the EM task**.
  - Reducing the number of input tokens save **computational resources**, quicken the inference time, and **save financial resources.**

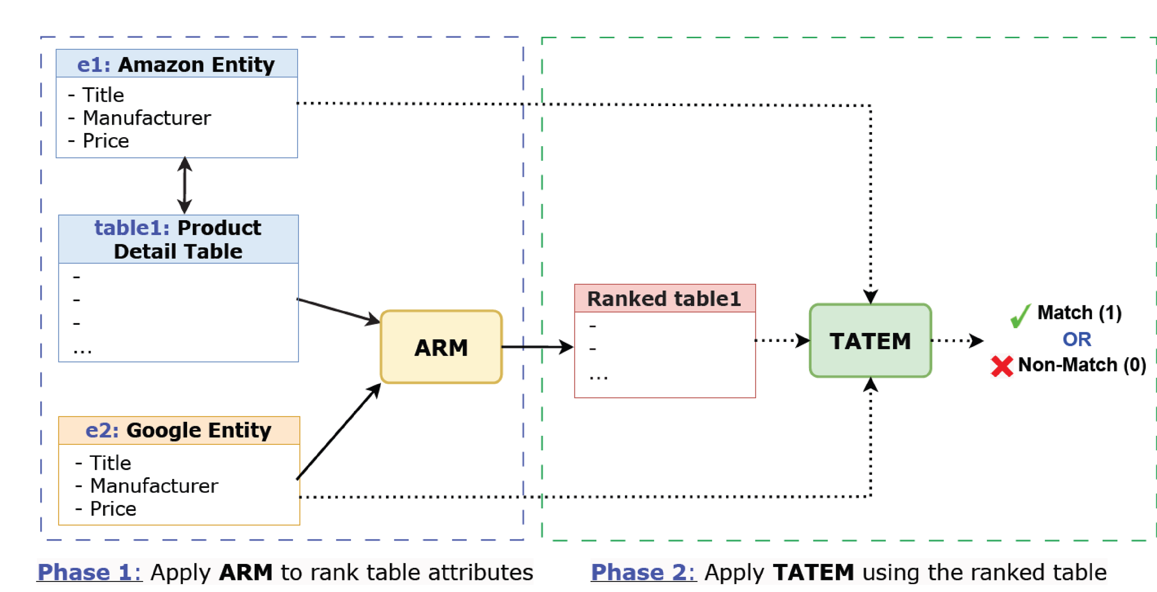

Figure 2: Our TATEM model coupled with ARM for PEM.

ARM calculates the relevance of Amazon detail table attributes, $attr\_i$, with respect to a Google product context, and it **returns the top n attributes** based on these estimates of relevance.

$$P(Relevence = 1 \mid attr_i, cntx) \triangleq \phi(\eta_{attr}(attr_i), \eta_{cntx}(cntx))$$

Our design choice for both encoders is Sentence-BERT (SBERT), and **we utilize cosine similarity** as the comparison function and title as the Google product context.

Experimental Results

Models	F1 Score	
	Amazon-Google [20]	**Walmart-Amazon** [20]
DM+ (2018)	70.7	73.6
DITTO (2020)	75.58	86.76
KAER (2023)	76.25	-
ROBEM (2022)	79.06	86.68
SupCon (2022)	79.28	-
GPT3(k=0) (2022)	54.3	60.6
GPT3(k=10) (2022)	63.5	87.0
	Amazon-Google-Tab (*ours*) [structured]	**Walmart-Amazon-Tab** (*ours*) [structured]
DITTO	80.56	86.85
ROBEM	78.50	85.74
SupCon	78.58	-
	Amazon-Google-Tab (*ours*)	**Walmart-Amazon-Tab** (*ours*)
DITTO-m	79.35	86.42
ROBEM-m	80.92	88.31
TATEM (*ours*)		
+ w/ *all tuples*	**82.2**	**90.56**
+ w/ ARM (n=1)	80.12	88.52
+ w/ ARM (n=3)	81.28	89.24
+ w/ ARM (n=5)	81.83	89.77

Table 3: Performance of TATEM compared to different baselines. All reported results for TATEM (*ours*) are statistically significant in paired t-test by taking DITTO (2020) as a reference with the confidence of 95% (p-value < 0.05).

- Results emphasizes the **advantages of directly serializing** semi-structured data, particularly for lengthy, complex product tables.
- TATEM, reaches **new SOTA results** (F1 score of 82.2 and 90.56 for Amazon-Google-Tab and Walmart-Amazon-Tab.