Segment-driven Structural Induction and Semantic Alignment for Heterogeneous Tabular Representation

Woojun Jung, Susik Yoon
Korea University, Seoul, South Korea

Paper Code Contact

What did we model?

Across heterogeneous tables, we distinguish attributes from their observed realization.

"Attributes"

An attribute is a canonical textual identifier shared across tables

A set of attributes forms a schema: a logical frame for describing an entity or relation.

"Segments"

A segment is a symbolic header-value realization of an attribute within a table

evidence! *-Realization-*

TITLE	DIRECTOR	GENRE
Star Wars	Spielberg	Sci-fi
Tenet	Nolan	Sci-fi
Interstellar	Nolan	Sci-fi

Across tables, these realizations provide evidence for what an attribute means in context.

Transferability–Specialization Trade-off for In-Domain Tables

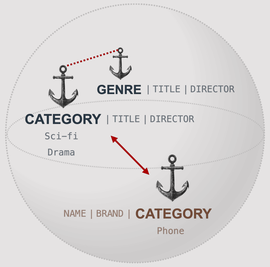
Movie domain

1. Popular Movies

TITLE	DIRECTOR	GENRE
Interstellar	Nolan	Sci-fi
Oppenheimer	Nolan	Drama

2. Filmography, Masterpiece

DIRECTOR	TITLE	CATEGORY
Nolan	Interstellar	Sci-fi
Cameron	Titanic	Drama



Product domain

3. Mobile Devices

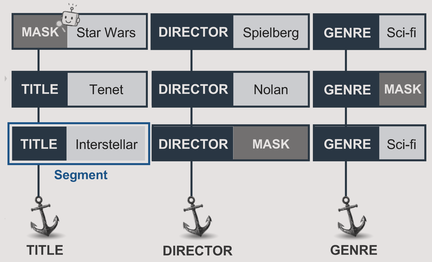
NAME	BRAND	CATEGORY
iPhone 17 Pro	Apple	Phone

Across heterogeneous in-domain tables, the same header can denote different concepts, while different headers can refer to the same concept—making header names alone insufficient for capturing table-specific meaning and motivating **NAVI**.

How we built NAVI —1

Masked Segment Modeling for Schema Context

(Structural Evidence)



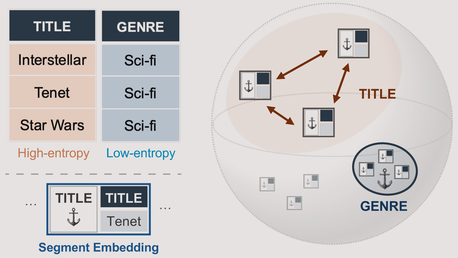
Optimize MLM objective based on field-level (header/value) and vanilla MLM masking.

How we built NAVI —2

Entropy-driven Segment Alignment for Column Context

(Distributional Evidence)

TITLE	GENRE
Interstellar	Sci-fi
Tenet	Sci-fi
Star Wars	Sci-fi
High-entropy	Low-entropy



Optimize InfoNCE loss on segment embeddings routed by normalized entropy of columns.

How did we evaluate?

Dataset. We evaluate NAVI on heterogeneous web tables from the Movie and Product domains in *WDC WebTables*. 8:1:1 train/val/test split and evaluate on held-out test rows.

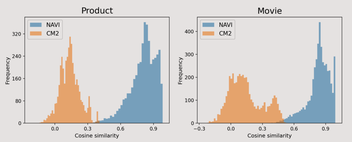
Baselines. We compare NAVI with BERT (generic LM), TAPAS (structure-aware), CM2 (feature-oriented), and HAETAE (domain-aware).

Criteria. We evaluate whether representations are:

Generative *Can they reconstruct masked headers and values?*	Coherent *Do equivalent attributes cluster together?*
Discriminative *Do they improve downstream prediction?*	Robust *Do they withstand schema perturbations?*

Can NAVI recover and use meaning? & Do semantics survive schema variations?

Model	Product		Movie	
	Header	Value	Header	Value
BERT	0.9400	0.7743	0.9233	0.6889
TAPAS	0.2996	0.2501	0.2619	0.2326
HAETAE	0.9260	0.7785	0.8982	0.6911
NAVI	**0.9995**	0.7933	**0.9998**	**0.7077**
w/o CAE	0.5996	0.5544	0.5485	0.4799
w/o MSM	0.9959	0.7769	0.9941	0.6792
w/o ESA	0.9993	**0.7977**	0.9994	0.7012

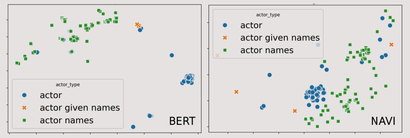


NAVI achieves near-perfect header recovery and the strongest value imputation overall (left); its masked fields also remain substantially closer to their original semantic representations than CM2's (right).

Category	Model	Product XGBoost	Product LR	Product TabPFN	Movie XGBoost	Movie LR	Movie TabPFN
LM Encoders	BERT	0.912 (± 0.013)	0.931 (± 0.015)	0.937 (± 0.020)	0.595 (± 0.036)	0.638 (± 0.043)	0.659 (± 0.038)
	TAPAS	0.911 (± 0.013)	0.925 (± 0.015)	0.930 (± 0.013)	**0.629 (± 0.035)**†	**0.669 (± 0.023)**†	**0.688 (± 0.017)**†
	CM2	0.814 (± 0.023)	0.846 (± 0.018)	0.842 (± 0.017)	0.379 (± 0.025)	0.375 (± 0.020)	0.402 (± 0.015)
	HAETAE	0.915 (± 0.015)	0.938 (± 0.011)	0.933 (± 0.013)	0.602 (± 0.050)	0.644 (± 0.041)	0.664 (± 0.016)
	NAVI	**0.928 (± 0.010)**	**0.939 (± 0.012)**	**0.946 (± 0.013)**†	0.615 (± 0.027)	0.667 (± 0.028)	0.679 (± 0.030)
	w/o CAE	0.725 (± 0.025)	0.854 (± 0.026)	0.801 (± 0.019)	0.265 (± 0.027)	0.381 (± 0.028)	0.303 (± 0.021)
	w/o MSM	0.845 (± 0.024)	0.903 (± 0.012)	0.894 (± 0.015)	0.467 (± 0.025)	0.560 (± 0.027)	0.522 (± 0.037)
	w/o ESA	0.909 (± 0.019)	0.930 (± 0.012)	0.934 (± 0.013)	0.623 (± 0.010)	0.667 (± 0.010)	0.674 (± 0.010)
FE Pipelines	Raw	0.887 (± 0.015)	0.829 (± 0.013)	0.910 (± 0.013)	0.457 (± 0.038)	0.420 (± 0.035)	0.478 (± 0.019)
	TableVec.	0.920 (± 0.017)	0.915 (± 0.014)	0.931 (± 0.019)	0.622 (± 0.014)	0.552 (± 0.029)	N/A
	NAVI$_{FE}$	**0.944 (± 0.017)**†	**0.942 (± 0.012)**†	**0.942 (± 0.015)**	**0.629 (± 0.027)**	**0.661 (± 0.028)**	**0.665 (± 0.031)**

These gains transfer to downstream learning: NAVI is strongest among other baselines across classifiers in Product domain ,while NAVI$_{FE}$ further improves conventional feature-engineering pipelines.

Model	Product NMI	Product B^3-F1	Movie NMI	Movie B^3-F1
BERT	0.8749	0.7317	0.8798	0.6969
TAPAS	0.8750	0.7239	0.8726	0.6759
HAETAE	0.8742	0.7268	0.8864	0.7276
NAVI	**0.9005**	**0.7920**	**0.9144**	**0.7996**



NAVI consolidates lexical variants into compact, well-separated attribute clusters, achieving the strongest NMI and B^3-F1 across both domains (left). Compared with BERT, it reduces inter-variant cosine and L2 distances from 0.36 to 0.30 and from 0.77 to 0.72, while increasing the silhouette score from 0.26 to 0.52—improving alignment without collapsing distinct attributes (illustrated in right).

	Product Default	Product Col. Perm.	Product Synonym	Product Typo	Movie Default	Movie Col. Perm.	Movie Synonym	Movie Typo
BERT	.9400	.8730 (-7.12%)	.9189 (-2.24%)	.8855 (-5.80%)	.9233	.8590(-6.96%)	.9044 (-2.05%)	.8762 (-5.10%)
TAPAS	.2996	.2681 (-10.51%)	.2457 (-17.99%)	.2539 (-15.25%)	.2619	.2331 (-11.00%)	.2472 (-5.61%)	.2328 (-11.11%)
HAETAE	.9260	.8619 (-6.92%)	.9146 (**-1.23%**)	.8800 (**-4.97%**)	.8982	.8189 (-8.83%)	.8889 (**-1.04%**)	.8504 (-5.32%)
NAVI	.9995	**.9997 (+0.02%)**	**.9760** (-2.35%)	**.9375** (-6.20%)	.9998	**.9997 (-0.01%)**	**.9870** (-1.28%)	**.9522 (-4.76%)**

NAVI remains highly stable under column permutations, synonym substitutions, and typos, indicating that attribute semantics are not tied to fragile schema surface forms.

Yes. NAVI learns domain-specialized attribute semantics through accurate header value reconstruction, coherent alignment of lexical variants, and stronger downstream prediction—while remaining stable under schema perturbations.