

Contrastive Entity Linkage: Mining Variational Attributes from Large Catalogs for Entity Linkage

AKBC 2020

Varun Embar, Bunyamin Sisman, Hao Wei, Xin Luna Dong,
Christos Faloutsos and Lise Getoor

Motivation



iPhone 11 Pro 64 GB



iPhone 11 Pro 256 GB

Are these two entities the **same** or **different**?

Motivation

Attributes



iPhone 11 Pro 64 GB

Same

Brand

Color

Generation

Different

Storage



iPhone 11 Pro 256 GB

Motivation



iPhone 11 Pro 64 GB



iPhone 11 Pro 128 GB

Variations

Base
Attributes

Same

Brand
Manufacturer
Model

Variational
Attributes

Different

Color
Storage

Motivation

	apple	11
	amazon	5
	bose	qc11

Catalog 1

	bose	qc11
	apple	11
	bose	qc11

Catalog 2

Entity
Linkage

Duplicates



Distinct



Variations



Contributions

[C1] Automatic variational attribute discovery

- Propose *contrast feature* that model variation attributes
- Novel scalable, unsupervised *VarSpot* algo to extract them

[C2] Three-way entity linkage

- Distinct, variation and duplicates
- Contrastive entity linkage framework

[C3] Effectiveness

- Empirical evaluation on three different domains
- Three different entity linkage frameworks

Related Work

	Duplicate Matching	Variation Matching	Variational Attribute Extraction
Entity Linkage Approaches[1]	✓		
GROUP Li et al. [2015] Recasens et al. [2011]		✓	
Attribute Extraction Techniques [2]			✓
Contrastive Entity Linkage	✓	✓	✓

[1] Christen et. al. 2012, Rahm, 2010, Halevy 2005, Machanavajjhala 2012 etc.

[2] Zheng 2018, Bizer 2017, Weld 2012, Hu 2011, Kannan 2011 etc.

Approach - VarSpot

	apple	11
	amazon	5
	bose	qcll

Catalog 1

	apple	11
	amazon	5
	bose	qcll

Catalog 1

Phase 1

Blocking
&
Linkage

Same Catalog

C1



See paper for more details

Approach - VarSpot

C1

Phase 2



Apple iPhone 11 Pro 64 GB






Apple iPhone 11 Pro 256 GB

Contrast
features

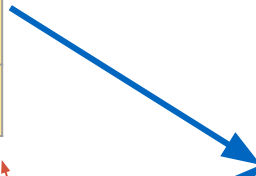
Approach - Contrastive entity linkage C2

	apple	11	white
	amazon	5	black
	bose	qcII	black

Catalog 1

	bose	qcII	rose
	apple	11	black
	bose	qcIII	black

Catalog 2



Duplicates



Distinct



Variations



Extracted contrast features

Evaluation

C3

Domains

- Software (Small-sized dataset)
- Groceries (Medium-sized dataset)
- Music (Large-sized dataset)

Entity linkage frameworks

- Magellan [Konda et. al. 2016]
- SILK [Isele et. al. 2010]
- Deepmatcher [Mudgal et. al. 2018]

Evaluation

C3

Variations identified by VarSpot algorithm

Groceries

Milk duds candy 1.85 ounce boxes pack of 24

Milk duds candy 5 ounce boxes pack of 3

Milk duds movie size 5 oz 12 count

Music

Groove is in the heart

Groove is in the heart club version

Groove is in the heart sampladelic remix

Software

Peachtree by sage premium accounting for nonprofits 2007

Peachtree by sage premium accounting 2007 accountants' edition

Peachtree by sage pro accounting 2007

Evaluation

C3

Top contrast features identified by VarSpot algorithm

Software	Groceries	Music
standard mac upgrade	pack of 6	remix
small box	pack of 2	mix
premium upsell mac	2 pack	radio edit
standard upsell mac	red	live
deluxe	strawberry	instrumental

Evaluation

C3

Magellan

Software		Without contrast features	CEL
Duplicates	F1	0.785	0.81
	APS	0.877	0.897
Variations	F1	0.677	0.695
	APS	0.761	0.777

CEL significantly outperform models without contrast features

More results in the paper

For more details visit our poster # fR44nF03Rb