Knowledge Cleaning

Why Knowledge Cleaning?

Tools & Home Improvement › Paint, Wall Treatments & Supplies › Wall Stickers & Murals


Roll over image to zoom in

alasijia White Summer Magnetic Mesh Net Anti Mosquito Insect Fly Bug Curtain Automatic Closing Door Screen Kitchen Curtain-90CMx210CM

Brand: alasijia

Currently unavailable.

We don't know when or if this item will be back in stock.

Color	90cmx210cm
Material	Plastic, Fabric
Brand	Alasijia
Surface Recommendation	Door

About this item

- Leave your door open and enjoy fresh cooler air,Completely prevent mosquitoes, spiders, moths, flies, bugs and other flying insects go into the room.
- perfect Bug & Mosquito Net For Door, Bring You Comfort, Free Your Hands To Entry, As Well As Ensure Your Little Baby And Pet Can Easily To Access.you Don't Have To Wake Up On A Good Weekend Morning To Open Doors For Pets And babies.
- Great natural insect protection for open balconies&patio doors,Foldable&easy to store,Fits over single doors, sliding doors&caravan doors,Essential accessory to any home during the summer months
- Material: Polyester fiber.lightweight mesh screen with almost no sound when switching, You won't be disturbed while sleeping or working。
- pay attention：please carefully measure your door frame size before purchase,The size of the panel needs to be 3cm wider than the door frame and 6cm high.

Color of kitchen curtain: “90CM X 210CM”?

Product information

Manufacturer	alasijia
ASIN	B07S4KX3PB
Best Sellers Rank	#2,747,737 in Tools & Home Improvement (See Top 100 in Tools & Home Improvement) #223,977 in Wall Stickers & Murals
Scent	90CMx210CM

Scent of kitchen curtain: “90CM X 210CM”?

Why Knowledge Cleaning?

Beauty & Personal Care › Skin Care › Body › Cleansers › Soaps

Van Der Hagen Glycerin Soap, Free, 3.75 Ounce

Visit the Van Der Hagen Store

25 ratings | 3 answered questions

Price: $8.07 ($2.15 / Ounce)

Get 5% back ($0.40 in rewards) on the amount charged to your Amazon Prime Rewards Visa Signature Card.

Scent: **Free**

$8.07 ($2.15 / Ounce)	**$8.07 ($2.15 / Ounce)**
$8.10 ($2.16 / Ounce)	$7.99 ($2.13 / Ounce)

Brand	Van Der Hagen
Scent	Free
Item Weight	3.75 Ounces
Item Dimensions LxWxH	4 x 3 x 1.5 inches
Color	Free

About this item

- Hypo allergenic
- 44% humectant moisturizers
- Exceptionally mild
- Non-commedoginic will not clog pores
- Rinses clean leaves skin soft and smooth

“Free” is neither understandable scent nor color

Section Structure

- Problem Definition

  What is needed beyond techniques for building generic KGs?
- Short answer -- key intuition

  What are key intuitions for building product KGs?
- Long answer -- details

  What are practical tips?
- Reflection/short-answer

  Can we apply the techniques to other domains?

What is Knowledge Cleaning?

- Problem definition
  - Given a fact t = {**e, a, v**}, where
    - **e**: the product entity
    - **a**: an attribute of the product e
    - **v**: the attribute value of e
  - Identify if **t** states the true fact about **e**

Unique Challenges in Product Knowledge Cleaning

- Product Knowledge Graph has
  - Large number of entity types and relations
  - Rich unstructured textual information for entities
  - A large portion of triples are <entity, relation, literal/num>

Short answer/solution

- The key of knowledge cleaning is to detect data inconsistency
  - Among the values of the same attribute
  - Among the values of different attributes
  - Among different data sources

Short answer/solution

- Syntactic feature

Product	Attribute	Value
Kitchen curtain 1	color	white
Kitchen curtain 2	color	grey
Kitchen curtain 3	color	white
…	…	…
Alasijia magnetic kitchen curtain	color	**90CM X 210CM**

Incorrect attribute values are in **RED**

Short answer/solution

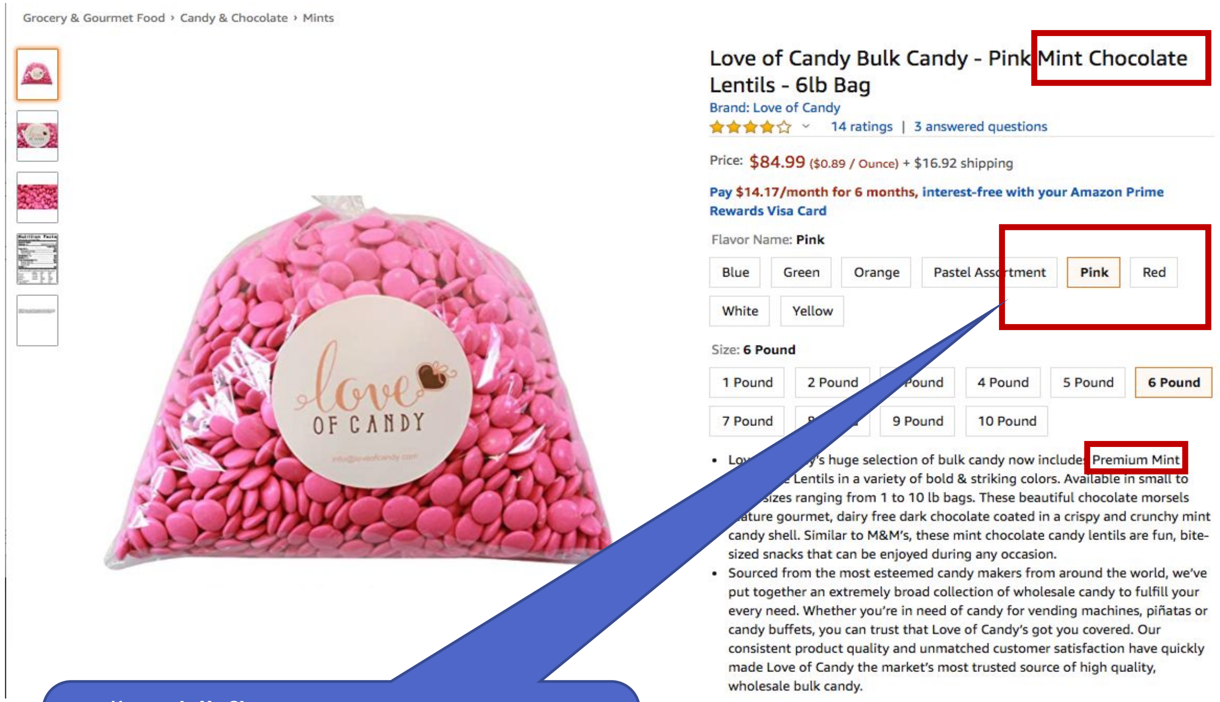
- Syntactic feature
- Rule/constraints

Product	Specialty	Sugar per serving
Syrup 1	Sugar free	0g
Syrup 2	Sugar free	0g
Syrup 3	N/A	15g
…	…	
Syrup 10	**Sugar free**	**15g**

Incorrect attribute values are in **RED**

Short answer/solution

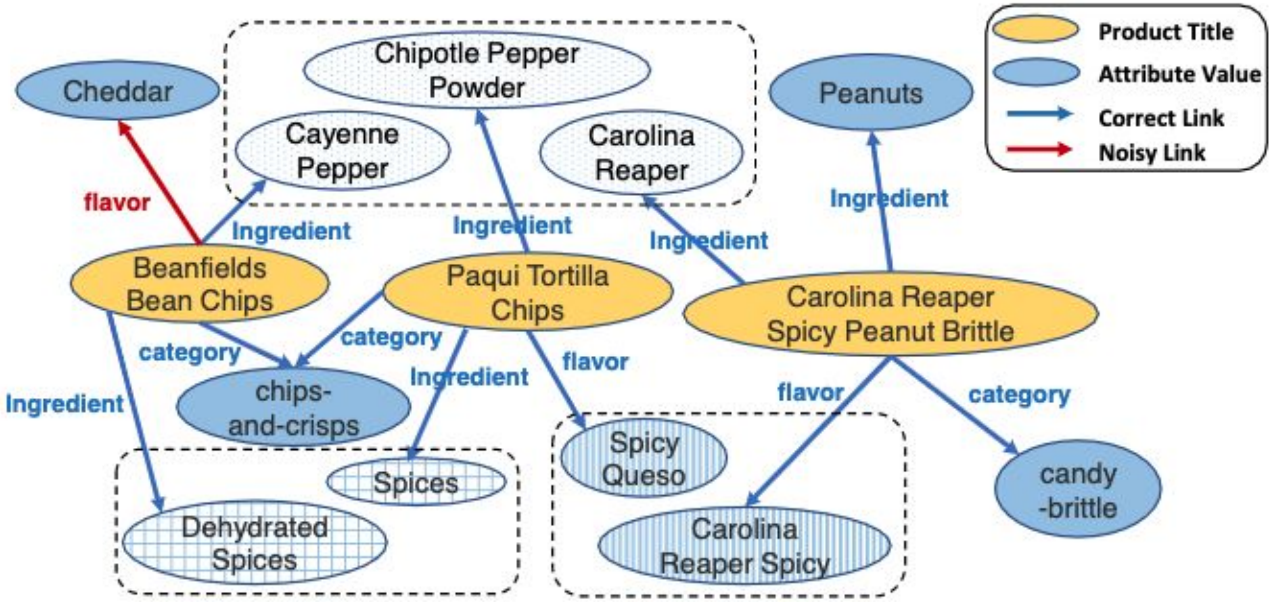
- Syntactic feature
- Rule/constraints
- Semantic understanding



“Pink” flavor is inconsistent with product’s title and bullet description

Short answer/solution

- Syntactic feature
- Rule/constraints
- Semantic understanding
- Graph embedding



Short answer/solution

- Syntactic feature
- Rule/constraints
- Semantic understanding
- Graph embedding
- Knowledge fusion

Source	Product	Material
Amazon	Alasijia magnetic kitchen curtain	Plastic
Walmart.com	Alasijia magnetic kitchen curtain	Plastic
Target.com	Alasijia magnetic kitchen curtain	Plastic
…	…	…
cookie.com	Alasijia magnetic kitchen curtain	**Linen**

Incorrect attribute values are in **RED**

Long answer: Syntactic features

- Recap: Intuition
    - Incorrect facts contain attribute values that violate the common syntactic patterns most values comply with

Product	Attribute	Value
Kitchen curtain 1	color	white
Kitchen curtain 2	color	grey
Kitchen curtain 3	color	white
…	…	…
Alasijia magnetic kitchen curtain	color	**90CM X 210CM**

Long answer: Syntactic features

- Auto-Detect [SIGMOD 2018]
  - Automatically detect incompatible values by leveraging an ensemble of judiciously selected generalization language

Sevilla - Jerez de la Frontera - Cádiz	1861
Córdoba - Málaga	1865.
Bobadilla - Granada	1874
Córdoba - Bélmez	1874
Osuna - La Roda	1877

(a) Extra dot

Polaco	15.04.1983	194	84
Vini	29.09.1982	N/A	N/A
Caio	30/11/1992	N/A	N/A
Jairo	17.02.1990	N/A	N/A
Michael	20.04.1983	N/A	N/A
Ricardinho	19.11.1975	192	94

(b) Mixed dates

2002 [12]	10.300 oz	899,500 oz
2005 [13]	25.272	2.174.620 oz
2006 [13]	49.354 oz	3.005.611 oz
2007 [13]	48.807 oz	3.165.408 oz
2008 [9]	47.755 oz	3.157.837 oz
2009[2]	0,9 million oz	818.050 oz

(c) Inconsistent weights

WARRIORS @ Sussex Thunder	13~28	—
WARRIORS @ Hampshire Thrashers	42~13	—
Essex Spartans @ WARRIORS	P~P	Postponed
WARRIORS @ Cambridgeshire Cats	36~44	—
East Kent Mavericks @ WARRIORS	12~18	—
WARRIORS @ East Kent Mavericks	15~17	—

(d) Score placeholder

No.	Title	Length
1.	"Cannibal vs. Cunning"	3:28
2.	"Lioness"	3:27
3.	"Self-Destruct & Die"	3:36
4.	"Narcotic"	3:00
5.	"In Coma"	4.00
6.	"Long Forgotten"	3:20

(e) Song lengths

47 806 (7,55%)	7
38 547 (6,09%)	6
26 824 (4,24%	4
21 604 (3,41%)	3
19 297 (3,05%)	3
16 861 (2,66%)	2

(f) Parenthesis

SK Rapid Wien	2:2
SV Mattersburg	2:4
SV Mattersburg	2 : 0
FC Wacker Innsbruck	2:4
SK Rapid Wien	1:1

(g) Scores

CAPT Thomas K. Chadwick	1992
CAPT Skip Blancett	1995
CAPT Leroy Gilbert	1998
CAPT Wilbur C. Douglass, III	2002
CAPT William F. Cuddy, Jr.	2006
CAPT Gary P. Weeden	June 11, 2010

(h) Mixed dates

Zhipeng Huang, Yeye He. Auto-Detect: Data-Driven Error Detection in Tables. SIGMOD 2018

Long answer: Syntactic features

- Auto-Detect [SIGMOD 2018]
  - Pattern Generalization

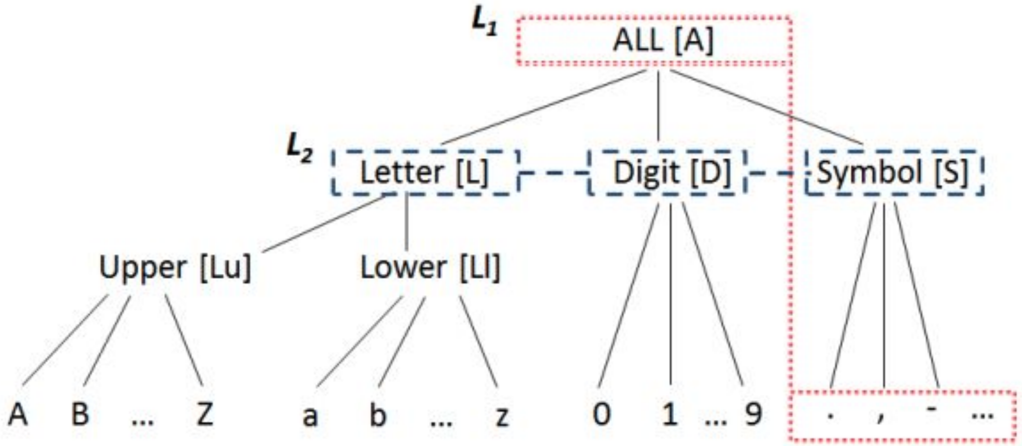


Figure 3: A generalization tree

EXAMPLE 2. L_1 and L_2 are two example generalization languages, each of which corresponds to a "cut" of the tree shown in Figure 3.

$$L_1(\alpha) = \begin{cases} \alpha, & \text{if } \alpha \text{ is a symbol} \\ \backslash\text{A}, & \text{otherwise} \end{cases} \tag{4}$$

$$L_2(\alpha) = \begin{cases} \backslash\text{L}, & \text{if } \alpha \in \{a, \cdots, z, A, \cdots, Z\} \\ \backslash\text{D}, & \text{if } \alpha \in \{0, \cdots, 9\} \\ \backslash\text{S}, & \text{if } \alpha \text{ is a symbol} \end{cases} \tag{5}$$

Given two values v_1 ="2011-01-01" and v_2 = "2011.01.02" in the same column, using L_1 we have

$$L_1(v_1) = \text{“}\backslash\text{A[4]-}\backslash\text{A[2]-}\backslash\text{A[2]”}$$
$$L_1(v_2) = \text{“}\backslash\text{A[4].}\backslash\text{A[2].}\backslash\text{A[2]”}$$

Zhipeng Huang, Yeye He. Auto-Detect: Data-Driven Error Detection in Tables. SIGMOD 2018

Long answer: Syntactic features

- Auto-Detect [SIGMOD 2018]
  - Distant supervision: generate training data

	$\mathbf{T}^+$					$\mathbf{T}^-$				
	t_1^+	t_2^+	t_3^+	t_4^+	t_5^+	t_6^-	t_7^-	t_8^-	t_9^-	t_{10}^-
L_1	0.5	0.5	-0.7	0.4	0.5	-0.5	0.9	-0.6	-0.7	0.2
L_2	0.5	0.5	0.4	-0.8	0.5	0.9	-0.6	0.2	-0.7	-0.7
L_3	0.4	0.5	0.5	0.6	0.5	-0.6	-0.6	-0.7	-0.5	0.9

Table 1: Generated training examples, where $t_i^+ = (u_i, v_i, +)$, $t_i^- = (u_i, v_i, -)$. Scores are produced based on NPMI after generalization in L_j is performed.

Zhipeng Huang, Yeye He. Auto-Detect: Data-Driven Error Detection in Tables. SIGMOD 2018

Long answer: Syntactic features

- Auto-Detect [SIGMOD 2018]
  - Aggregate predictions from languages
    - Dynamic-threshold aggregation: dynamically determine a separate threshold for each language and predict all cases below threshold as incompatible
    - Static-threshold aggregation: optimize the union of predictions to maximize the recall while maintaining a precision P
    - Greedy algorithm
      - Iteratively find a language L* from the candidate set LC, whose addition into the current selected set of candidate language G, will result in the biggest incremental gain
      - Iteratively expand the candidate set G using L* , until no further candidates can be added without violating the memory constraint

Zhipeng Huang, Yeye He. Auto-Detect: Data-Driven Error Detection in Tables. SIGMOD 2018

Long answer: Syntactic features

- Auto-Detect [SIGMOD 2018]

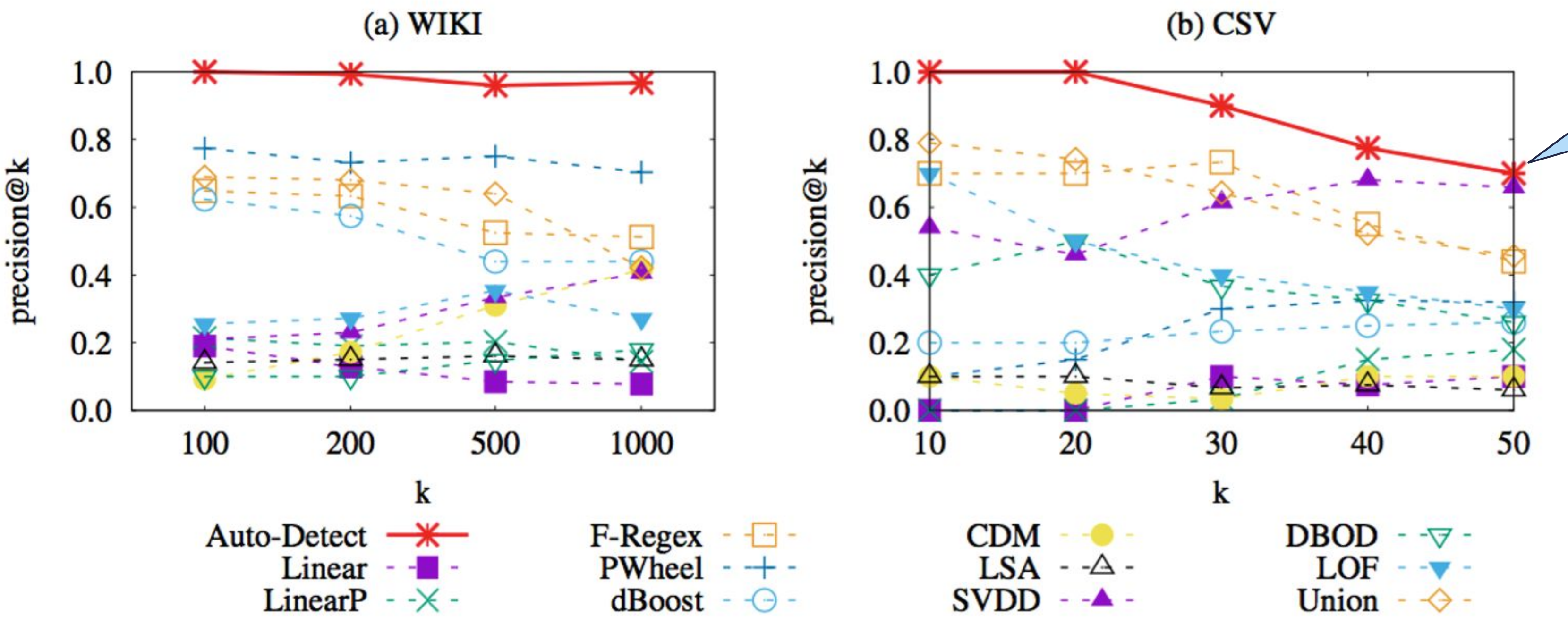


Figure 4: Quality results using manually labeled ground truth

Auto-Detect can find errors with high precision

Zhipeng Huang, Yeye He. Auto-Detect: Data-Driven Error Detection in Tables. SIGMOD 2018

Long answer: Rule/constraints

- Recap: Intuition
  - Discover declarative rules over the knowledge base and identify incorrect facts by finding data contradictions

Product	Specialty	Sugar per serving
Syrup 1	Sugar free	0g
Syrup 2	Sugar free	0g
Syrup 3	N/A	15g
…	…	
Syrup 10	**Sugar free**	**15g**

Long answer: Rule/constraints

- RuDik [ICDE 2018]
  - Discover both positive and negative rules over noisy and incomplete KBs
  - Generate positive and negative examples being aware of missing data and inconsistencies in KB
  - Incrementally materializes the KB as a graph, discover rules by navigating only the paths that potentially lead to the best rules

Stefano Ortona, Vamsi Meduri, Paolo Papotti. RuDik: Rule Discovery in Knowledge Bases. ICDE 2018

Long answer: Rule/constraints

- Example generation
- Incremental rule miner
- Rules execution

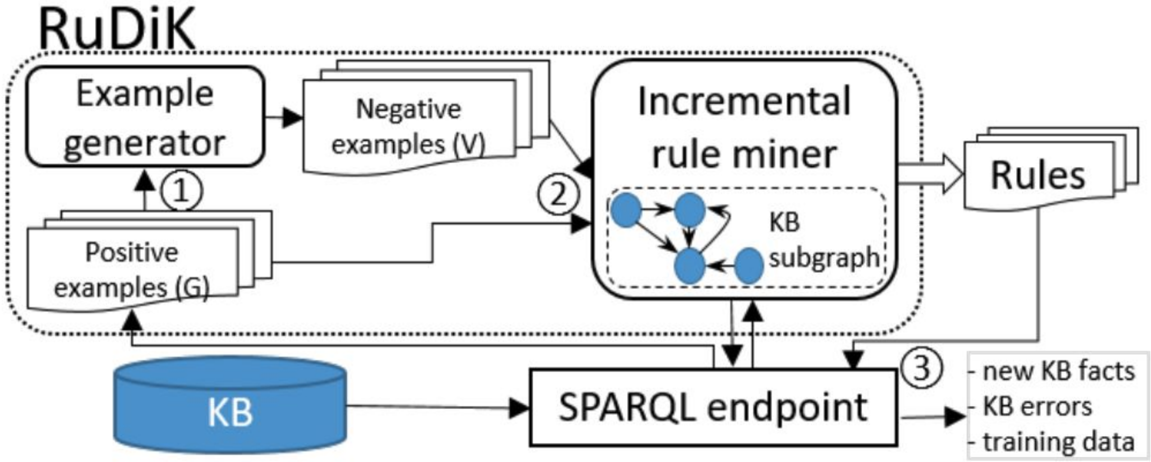


Figure 1: RuDiK architecture.

Stefano Ortona, Vamsi Meduri, Paolo Papotti. RuDik: Rule Discovery in Knowledge Bases. ICDE 2018

Long answer: Rule/constraints

TABLE II. RuDiK Positive Rules Accuracy.

KB	*Avg. RunTime*	*Avg. Precision over Predicates with Rules (All)*	*# Labeled Triples*
DBPedia	35min	**87.86%** (63.99%)	139
Yago 3	59min	**79.17%** (62.86%)	150
Wikidata	141min	**85.71%** (73.33%)	180

RuDIK showed very promising precision and rule discovery solution is scalable

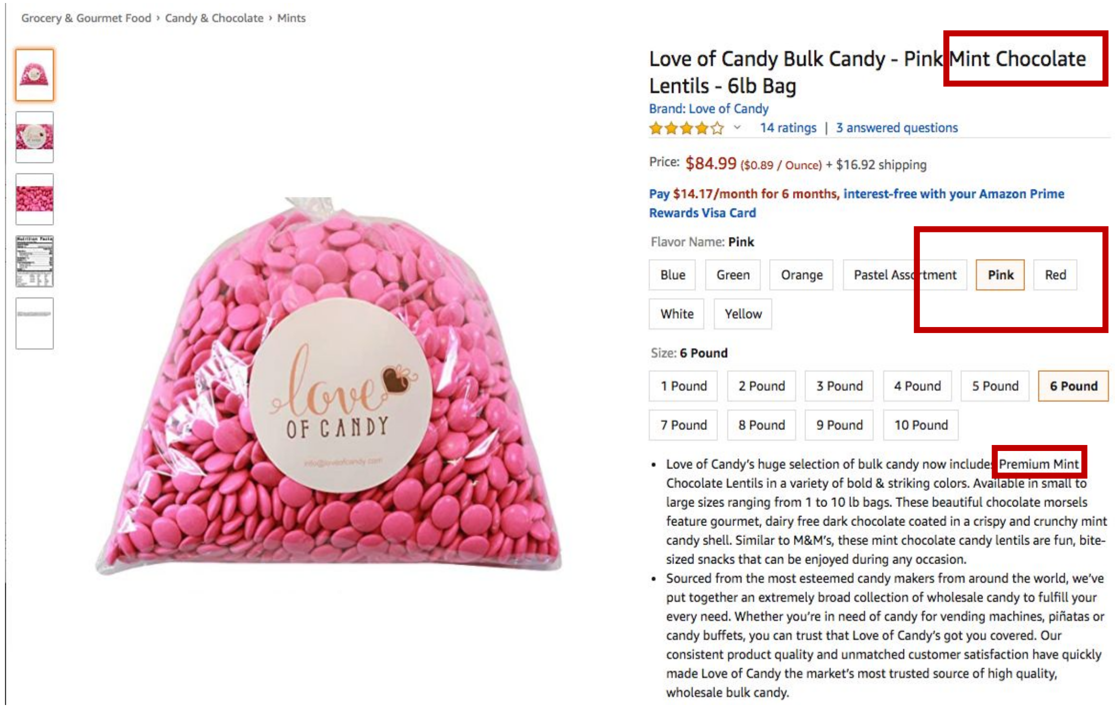
TABLE VI. Total Run Time Comparison.

KB	*#Predicates*	*AMIE*	RuDiK	*Types*
Yago 2	20	30s	18m,15s	12s
Yago 2s	26 (38)	> 8h	47m,10s	11s
DBPedia 2.0	904 (10342)	> 10h	7h,12m	77s
DBPedia 3.8	237 (649)	> 15h	8h,10m	37s
Wikidata	118 (430)	> 25h	8h,2m	11s
Yago 3	72	-	2h,35m	128s

Stefano Ortona, Vamsi Meduri, Paolo Papotti. RuDik: Rule Discovery in Knowledge Bases. ICDE 2018

Long answer: Semantic understanding

- Recap: Intuition
  - In retail domain, unstructured text includes rich information of product features, such as title, description, bullet points, etc.
  - The correctness of a fact can be validated by checking the consistency between the fact and the unstructured texts

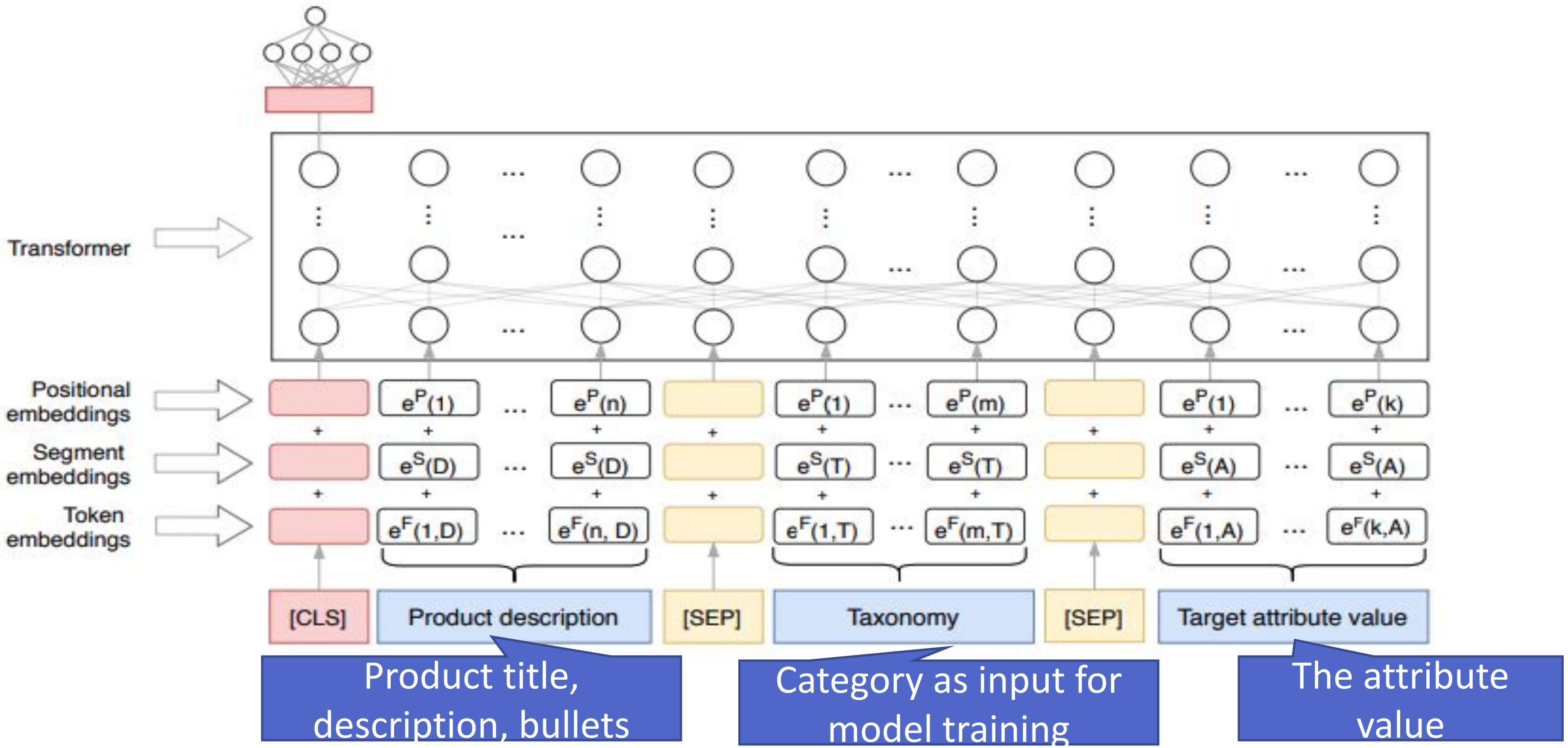


Long answer: Semantic understanding

- Auto-Know [KDD 2020]
  - Transformer-based model jointly processing signals from product profile, product taxonomy via multi-head attention to decide if an attribute value is correct
  - Use Amazon Catalog data for distant supervision
  - Use Snorkel to generate weak labels for training

Xin Luna Dong et al. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. KDD 2020

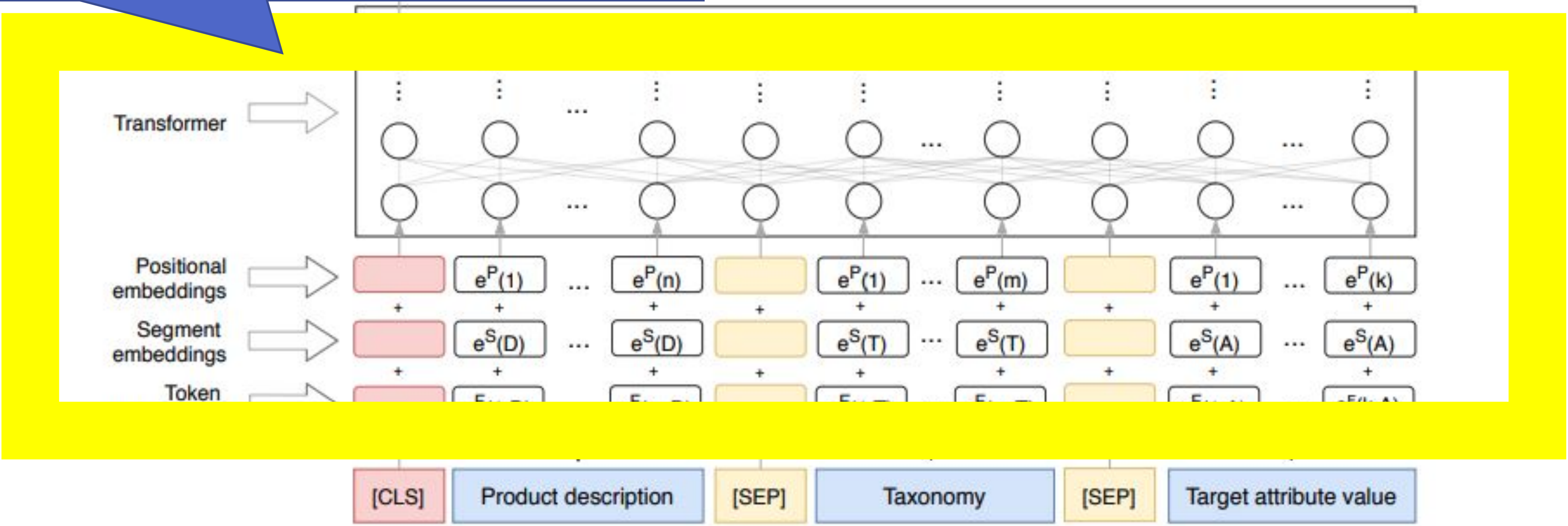
Long answer: Semantic understanding



Xin Luna Dong et al. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. KDD 2020

Long answer: Semantic understanding

Learn the semantic consistency between product profile, taxonomy and attribute value with **Transformer model**



Xin Luna Dong et al. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. KDD 2020

Long answer: Semantic understanding

- Experiment
  - Evaluated on 223 product categories

Model	PRAUC	R@.7P	R@.8P	R@.9P	R@.95P
Anomaly Detection [18]	32.0	2.4	1.3	1.3	1.3
AK-Cleaning	**56.1**	**59.6**	**39.8**	**26.0**	**20.7**
w/o. Taxonomy	52.6	52.6	36.2	22.4	3.0

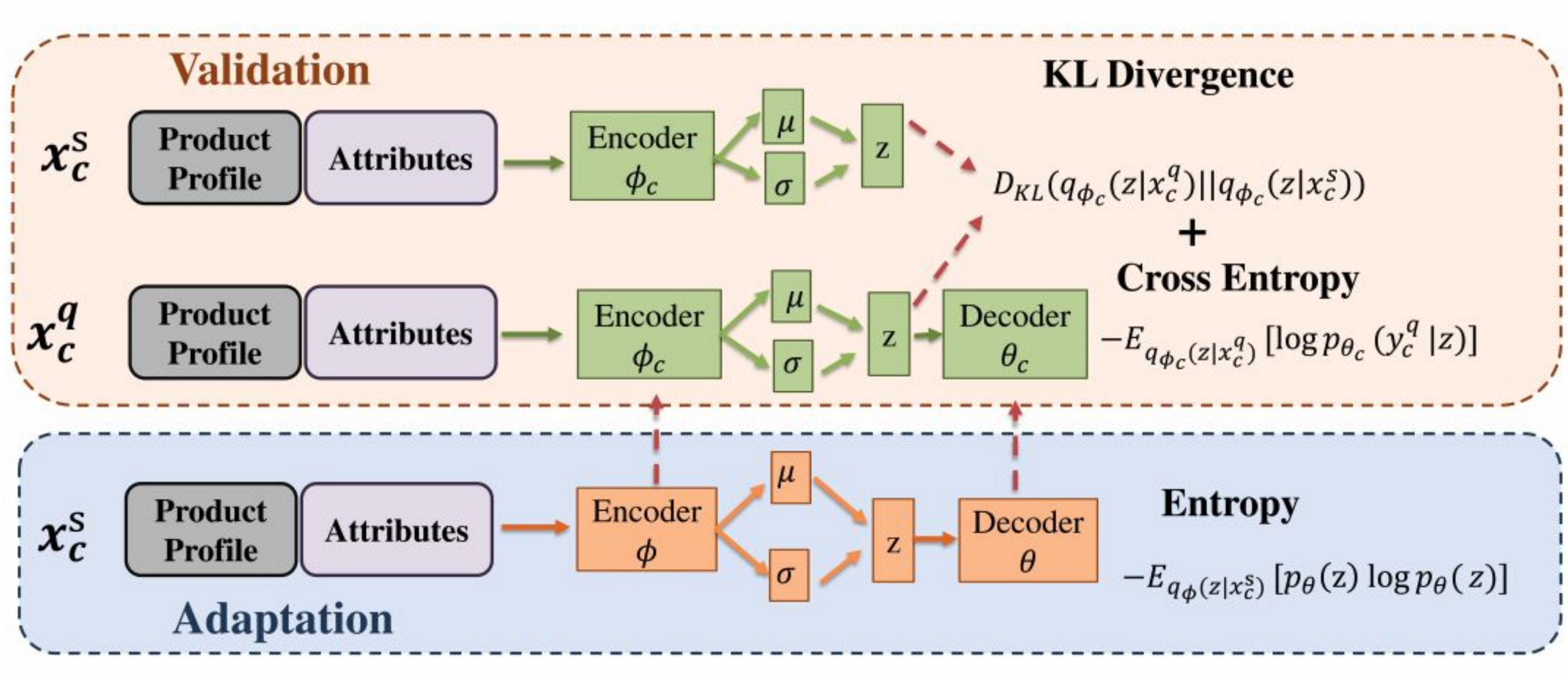
Xin Luna Dong et al. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. KDD 2020

Long answer: Semantic understanding

- MetaBridge [KDD 2020]
  - Few-shot learning setting
  - Integrate meta learning to make best use of labeled data from a small number of categories and ensure distribution consistency between unlabeled and labeled data and prevent overfitting
  - Combines meta learning and latent variable in a joint model to enhance the ability of capturing category uncertainty and preventing overfitting via effective sampling

Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, Jing Gao. Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data. KDD 2020

Long answer: Semantic understanding



Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, Jing Gao. Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data. KDD 2020

Long answer: Semantic understanding

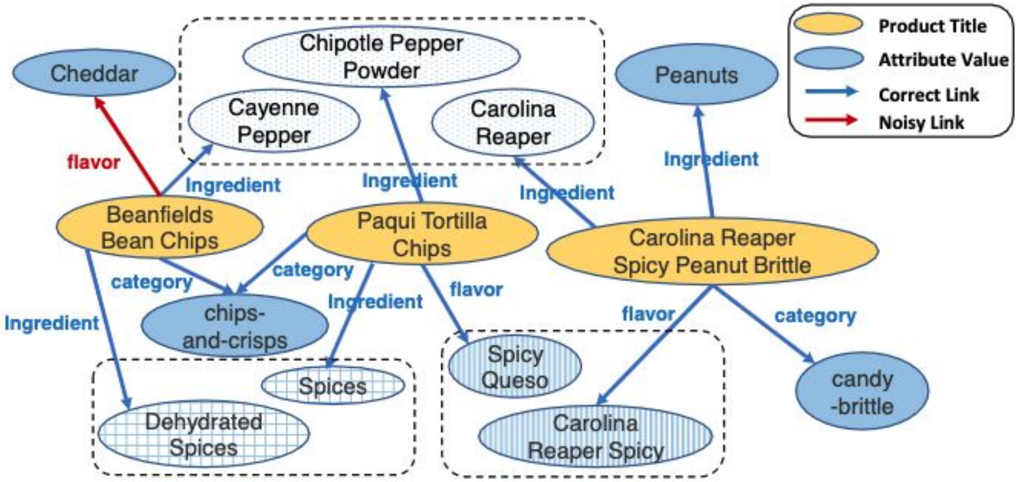
Setting	Method	Flavor		Ingredient	
		PRAUC	R@P=0.9	PRAUC	R@P=0.9
Supervised	RF	0.6986	4.43	0.4683	14.69
Fine-tune	BERT	0.7599	27.76	0.5292	17.00
Meta-Learning	MAML	0.7486	22.62	0.5289	22.48
Meta-Learning	MetaBridge	0.7852	30.77	0.5658	27.00

MetaBridge makes best use of training labels and outperforms supervised, fine-tuning methods

Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, Jing Gao. Automatic Validation of Textual Attribute Values in ECommerce Catalog by Learning with Limited Labeled Data. KDD 2020

Long answer: Graph embedding

- Recap: Intuition
  - Organize the products and all facts in a knowledge graph to explicitly reveal the correlation among different attributes
  - Learn KG embedding to capture the network structure
  - Incorrect facts usually contradict the global network structure



Long answer: Graph embedding

- Trans-E [NIPS 2013]
  - Treat relations as the translation operations between vectors corresponding to entities
  - Learn embeddings by minimizing a margin-based ranking criterion over the training set
  - Corrupt triples by replacing training triples with either head or tail replaced by a random entity

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko.
Translating Embeddings for Modeling Multi-relational Data. NeurIPs 2013

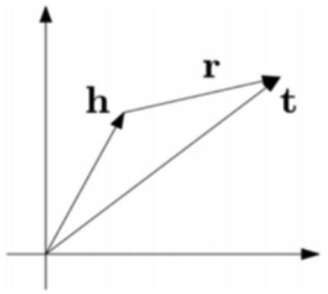
Long answer: Graph embedding

- Trans-E [NIPS 2013]
  - The score function of (h, r, t)

$$f_r(h,t) = -\ \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2}$$

  - Loss function

$$L = \sum_{\underbrace{(h,r,t)\in\triangle}_{\text{Positive triple set}}} \ \sum_{\underbrace{(h',r,t')\in\triangle'}_{\text{Negative triple set}}} \max(0, f_r(h,t) + \underbrace{M_{opt}}_{\text{Optimal Margin}} - f_r(h',t'))$$



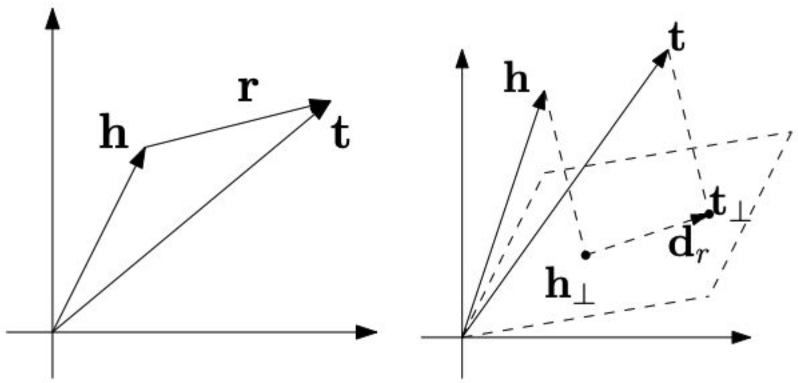
h + r = t
China + Capital = Beijing
France + Capital = Paris

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. NeurIPs 2013

Long answer: Graph embedding

- Trans-H [AAAI 2014]
  - Interprets a relation as a translating operation on a hyperplane
  - Each relation is characterized by two vectors
    - Norm vector of the hyperplane
    - Translation vector on the hyperplane
  - Score function of (h, r, t)

$$\|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r)\|_2^2$$
$$\mathbf{w}_r, \mathbf{d}_r \in \mathbb{R}^k$$



(a) TransE (b) TransH

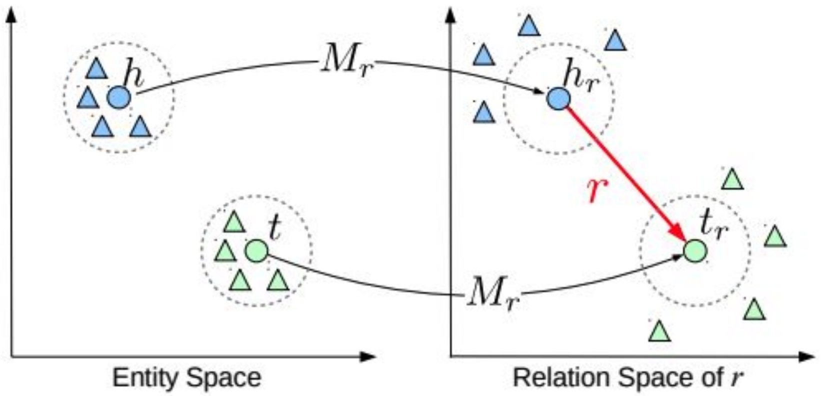
Zhen Wang, Jianwen Zhang, Jianlin Feng, Zheng Chen: Knowledge Graph Embedding by Translating on Hyperplanes. AAAI 2014

Long answer: Graph embedding

- Trans-R [AAAI 2015]
  - For each triple (h, r, t), entities in the entity space are first projected into r-relation space as hr and tr with operation Mr, then h_r + r = t_r
  - Scoring function of (h, r, t)

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r, \quad \mathbf{t}_r = \mathbf{t}\mathbf{M}_r.$$

$$f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$$



Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu: Learning Entity and Relation Embeddings for Knowledge Graph Completion. In AAAI, 2015

Long answer: Graph embedding

Data Sets	WN 11	FB13	FB15K
TransE	75.9	70.9	79.6
TransH	77.7	**76.5**	79.0
TransR	**85.5**	74.7	**81.7**

Knowledge Graph Embedding methods showed promising precision in detecting data errors

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, Xuan Zhu: Learning Entity and Relation Embeddings for Knowledge Graph Completion. In AAAI, 2015

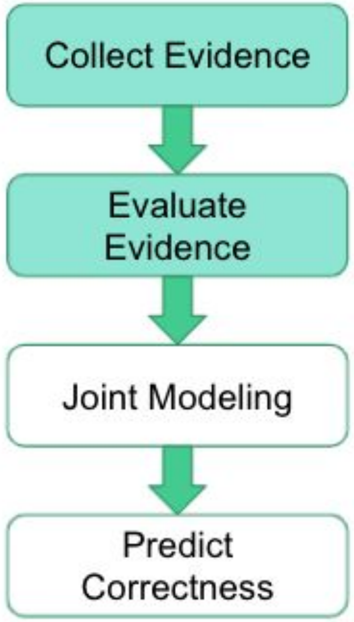
Long answer: Knowledge fusion

- Recap: Intuition
  - Data sources are of different quality and we trust data from accurate sources more

Source	Product	Material
Amazon	Alasijia magnetic kitchen curtain	Plastic
Walmart.com	Alasijia magnetic kitchen curtain	Plastic
Target.com	Alasijia magnetic kitchen curtain	Plastic
…	…	…
cookie.com	Alasijia magnetic kitchen curtain	**Linen**

Long answer: Knowledge fusion

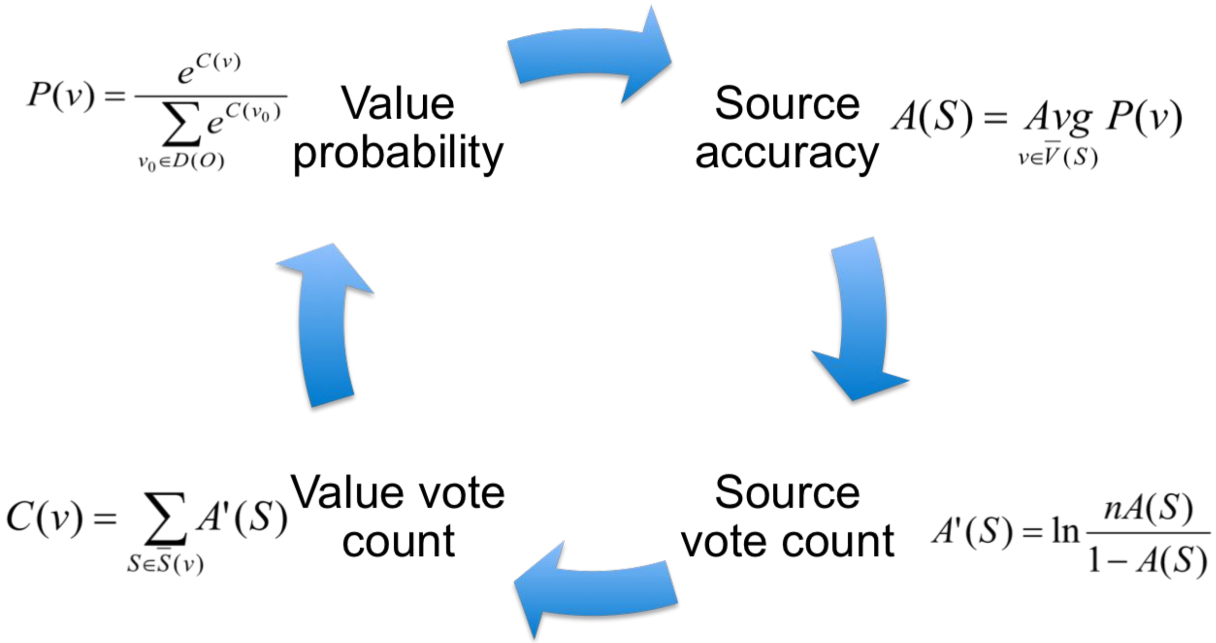
- ACCU [VLDB 2013]
  - Step 1/4: Gather evidence
    - Given a data item D (e.g. Obama's birthplace),
    - Dom(D) = $\{v_0, v_1, \ldots, v_n\}$
    - Φ: s_1 provides v_0 for D, s_2 does not provide any value for D
  - Step 2/4: Evaluate evidence
    - Objective evidence
    - Value distribution, similarity and formatting

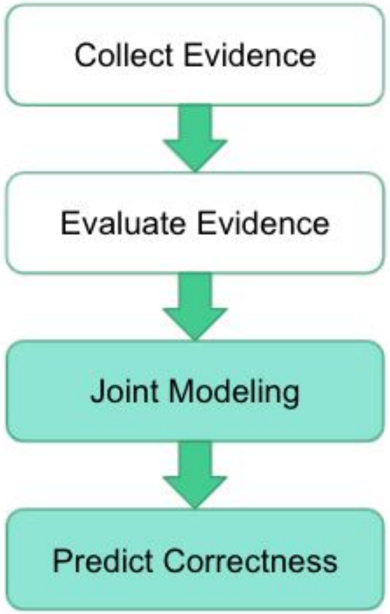


Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava: Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Long answer: Knowledge fusion

- ACCU [VLDB 2013]
  - Step 3/4: Prediction





Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava: Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Long answer: Knowledge fusion

- ACCU [VLDB 2013]
  - False value distribution
  - Value similarity
  - Value format
  - Trustworthiness on attribute level

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava: Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Long answer: Knowledge fusion

- ACCU [VLDB 2013]

Category	Method	Stock				Flight			
		prec w. trust	prec w/o. trust	Trust dev	Trust diff	prec w. trust	prec w/o. trust	Trust dev	Trust diff
Baseline	Vote	-	.908	-	-	-	.864	-	-
Web-link based	HUB	.913	.907	.11	.08	.939	.857	.2	.14
	AVGLOG	.910	.899	.17	-.13	.919	.839	.24	.001
	INVEST	.924	.764	.39	-.31	.945	.754	.29	-.12
	POOLEDINVEST	.924	.856	1.29	0.29	.945	***.921***	17.26	7.45
IR based	2-ESTIMATES	.910	.903	.15	-.14	.87	.754	.46	-.35
	3-ESTIMATES	.910	.905	.16	-.15	.87	.708	.95	-.94
	COSINE	.910	.900	.21	-.17	.87	.791	.48	-.41
Bayesian based	TRUTHFINDER	.923	.911	.15	.12	***.957***	.793	.25	.16
	ACCUPR	.910	.899	.14	-.11	.91	.868	.16	-.06
	POPACCU	.909	.892	.14	-.11	***.958***	***.925***	.17	-.11
	ACCUSIM	.918	***.913***	.17	-.16	.903	.844	.2	-.09
	ACCUFORMAT	.918	.911	.17	-.16	.903	.844	.2	-.09
	ACCUSIMATTR	***.950***	***.929***	.17	-.16	.952	.833	.19	-.08
	ACCUFORMATATTR	***.948***	**.930**	.17	-.16	.952	.833	.19	-.08
Copying affected	ACCUCOPY	**.958**	.892	.28	-.11	**.960**	**.943**	.16	-.14

Leverage source trustworthiness significantly improve the fact checking accuracy

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava: Truth finding on the Deep Web: Is the problem solved? In VLDB, 2013

Reflections/short-answers

- Knowledge cleaning is essentially to detect data inconsistency
    - Syntactic features
    - Learn rules and constraints
    - Semantic understanding
    - Graph structure
    - Multi-source integration
- All methods complement each other and effective ensemble them can maximize the final performance
- All these techniques are generic and applicable to KGs in other domains

Questions?

Break