Data Integration using Deep Learning Techniques

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Prof. Dr. Christian Bizer

Inspiration Session

May 4, 2022

Hello

- **Prof. Dr. Christian Bizer**
- Chair of Information Systems V:
  Web-based Systems

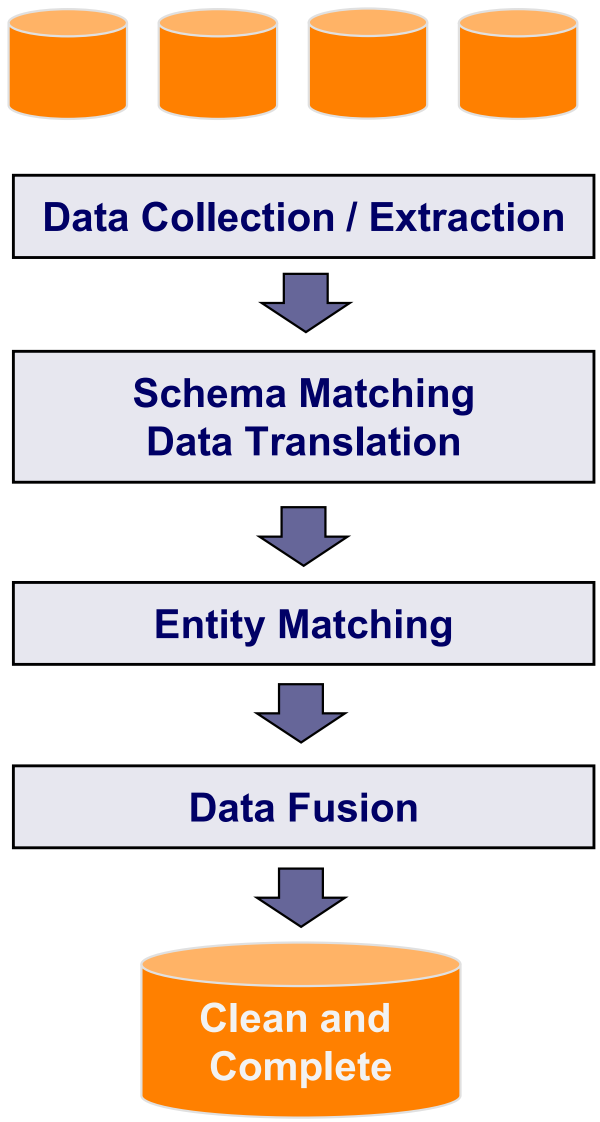
- Research Areas:
  - Large-scale data integration
  - Identity resolution, schema matching, data fusion
  - Information extraction from semi-structured sources
  - Knowledge base construction
  - Analysis of the adoption of semantic web technologies
- eMail: christian.bizer@uni-mannheim.de

Data and Web Science Group

- **Artificial Intelligence** (Prof. Heiner Stuckenschmidt)
- **Data Analysis** (Prof. Rainer Gemulla)
- **Data Science** (Prof. Heiko Paulheim)
- **Natural Language Processing** (Prof. Simone Ponzetto)
- **Process Analytics** (Prof. Han van der Aa)
- **Web-based Systems** (Prof. Christian Bizer)

The Data Integration Process

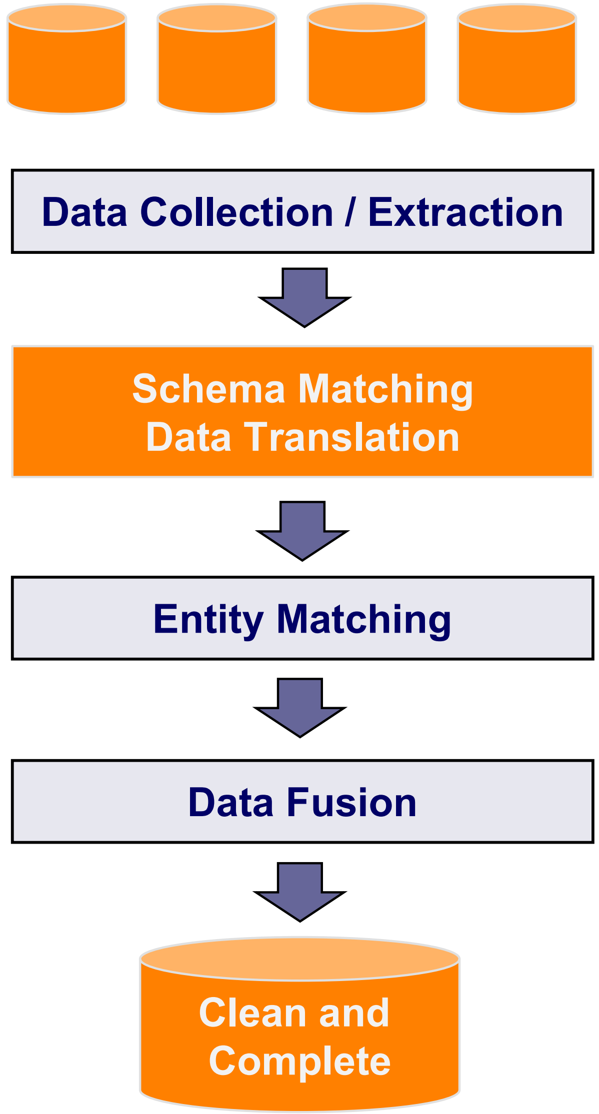


Schema Matching

Goal: Find correspondences between schema elements.

Name	Type	Memory	Description
Samsung Galaxy	S21 Blue	64 GB RAM	This great mobile …
Samsung Galaxy	S20	32 GB RAM	Last year's model …

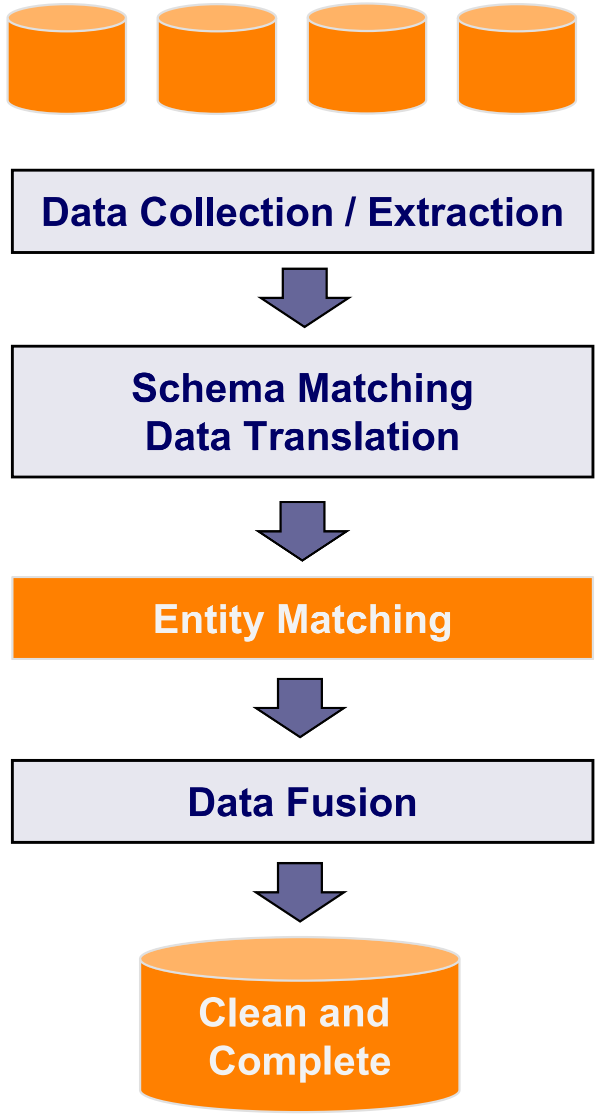
Brand	Product	Model No.	Mem in GB	Form Factor
Samsung	Galaxy	S21	64	Slate
Samsung	Galaxy	S20	32	Flat/Round



Entity Matching

Goal: Find all records that refer to the same real-world entity.

Brand	Product	Model No.	RAM	Color	Release
Samsung	Galaxy	S21	64	Blue	2021/1/29
Samung	Gal.	S 21 TGB12	64 GB	blau	Feb. 2021
	Galaxy S20 Blue TGB12 64GB		64000		2020/1/29



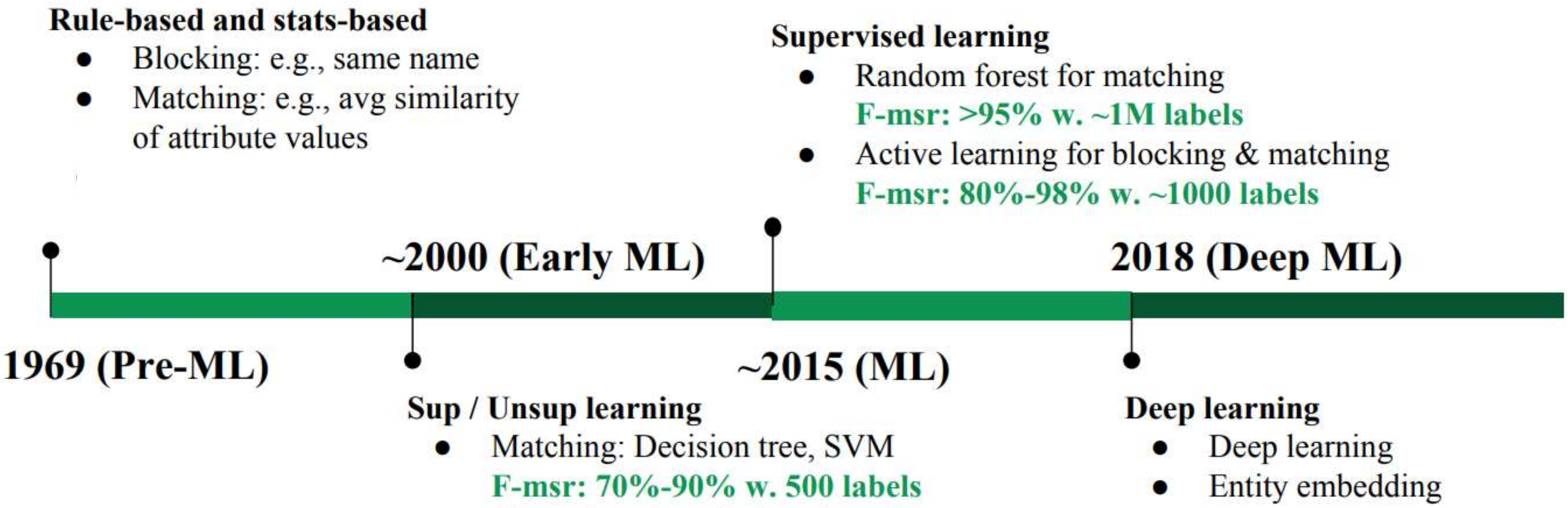
Outline

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. Entity Matching using Deep Learning Techniques
   – DeepMatcher
   – DITTO
   – JointBERT
   – Contrastive Pretraining
   – Summary
2. Schema Matching using Deep Learning Techniques
   – TURL
   – DoDuo
   – Summary
3. Conclusions

1. Entity Matching

50 Years of Entity Matching



Luna Dong: **ML for Entity Linkage. Data Integration and Machine Learning: A Natural Synergy**. Tutorial at SIGMOD 2018. https://thodrek.github.io/di-ml/sigmod2018/sigmod2018.html

Benchmark Datasets

Type	Dataset	Topic	# Pairs	# Matches	# Attributes
Structured	iTunes-Amazon	Music	539	132	8
	DBLP-ACM	Bibliographic	12,363	2,220	4
	DBLP-Scholar	Bibliographic	28,707	5,347	4
	Walmart-Amazon	Products	10,242	962	5
Textual	Abt-Buy	Products	9,575	1,028	3
	Amazon-Google	Products	11,460	1,167	3

https://github.com/anhaidgroup/deepmatcher/blob/master/**Datasets.md**

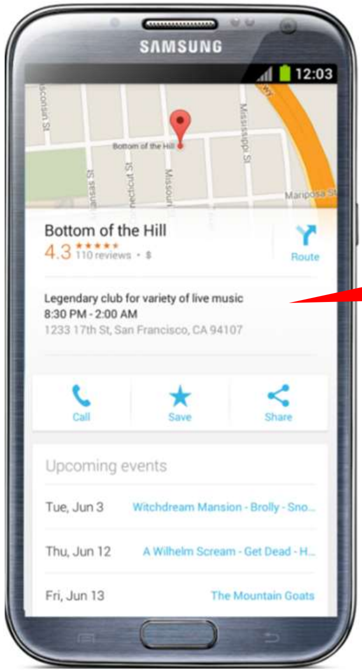
Anna Primpeli and Christian Bizer: **Profiling Entity Matching Benchmark Tasks**. CIKM 2020.

The Web as a Source of Training Data

- ask site owners since 2011 to annotate data within pages for enriching search results
- 675 Types: Event, Place, Local Business, **Product**, Review, Person
- Encoding: Microdata, RDFa, JSON-LD

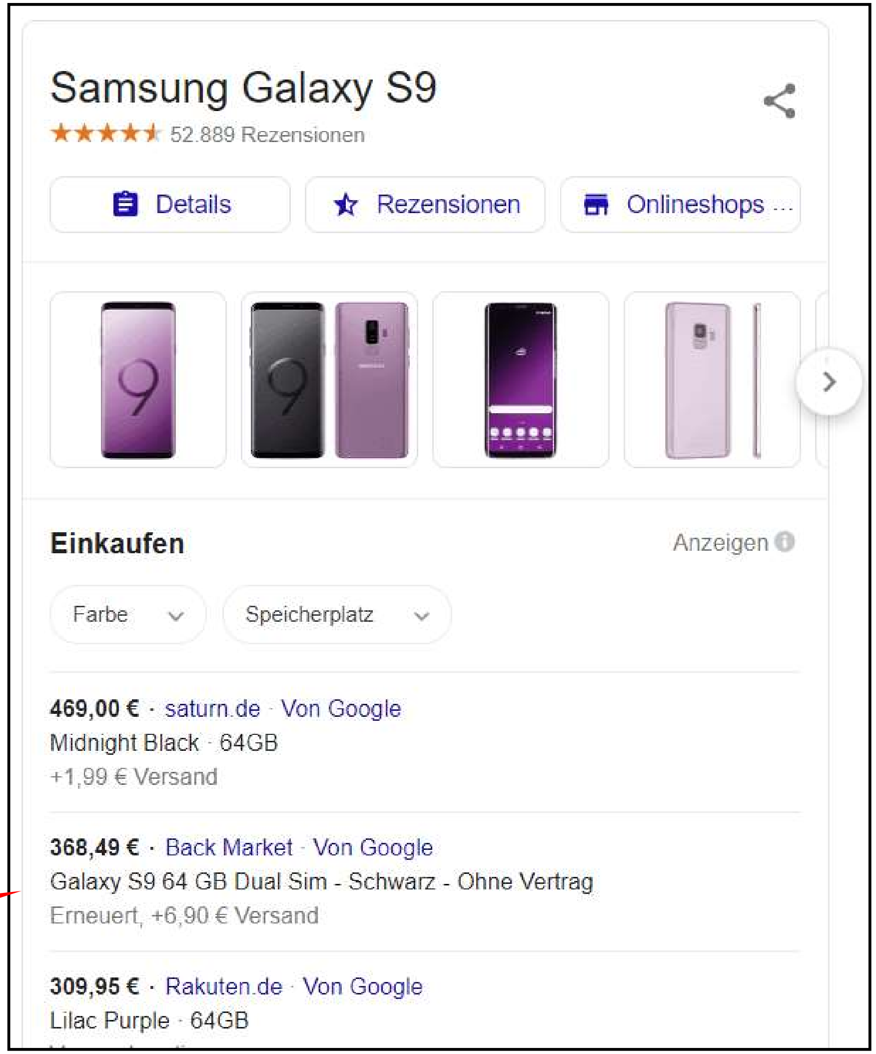
Usage of Schema.org Data @ Google

Data snippets within search results



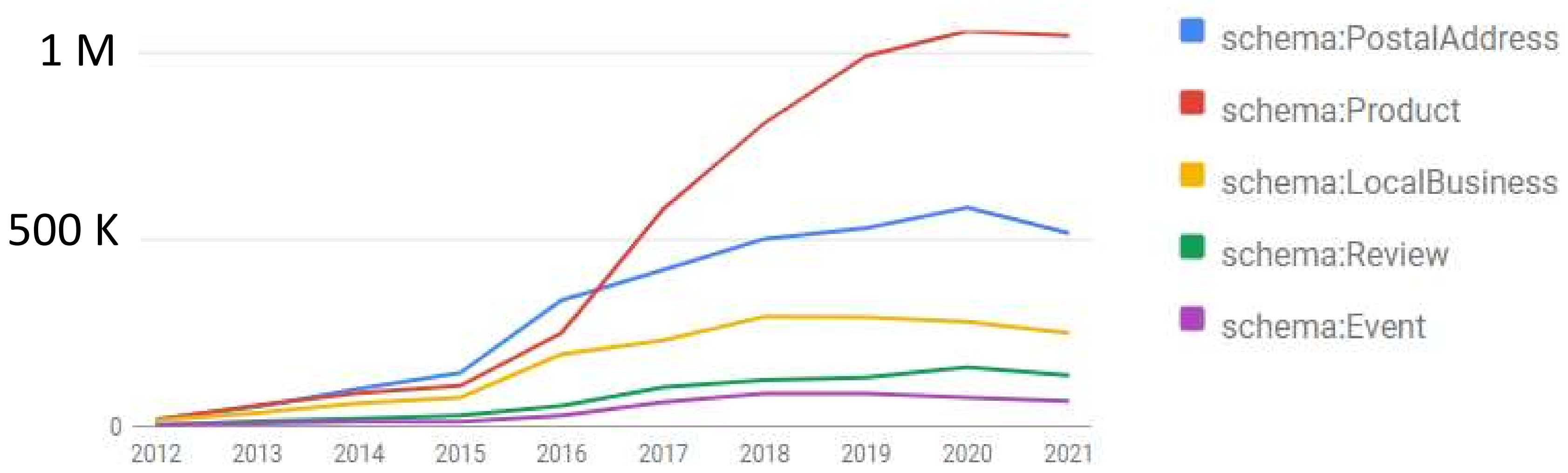
Local businesses on maps

Product offers



The Web Data Commons Project

- extracts all Microdata, RDFa, JSON-LD data from the Common Crawl
  - 3.1 billion HTML pages from 35 million websites (Oct. 2021)
  - 1.5 billion pages contain annotations (47.4%)
- Number of websites using specific schema.org types



http://webdatacommons.org/**structureddata**/

Example:
Schema.org Product Annotations within HTML

```
<div  itemtype="http://schema.org/Product">
     <h1 itemprop="name">Coleman Signature Instant Dome 7 Green</h1>
    Item # <span itemprop="gtin12">200734246807</span>
    <img itemprop="image" src="../20073424680.jpg">
   <p itemprop="description">The Coleman Signature 7-Person Instant Dome
        tent makes camping easy so you can enjoy every moment of your
        outdoor adventure. In about a minute you can have ...</p>
   <span itemprop="review" itemtype="http://schema.org/Review">
          <span itemprop="ratingValue">5</span>
          <span itemprop="reviewBody">Great waterproof tent!</span>
     </span>
</div>
```

WDC Product Data Corpus and Gold Standard for Large-Scale Product Matching

- 26 million product offers from 79.000 websites
  - offers are clustered by product IDs such as GTIN, MPN
- Training and test sets consisting of pairs of offers
  - different sizes, different product categories

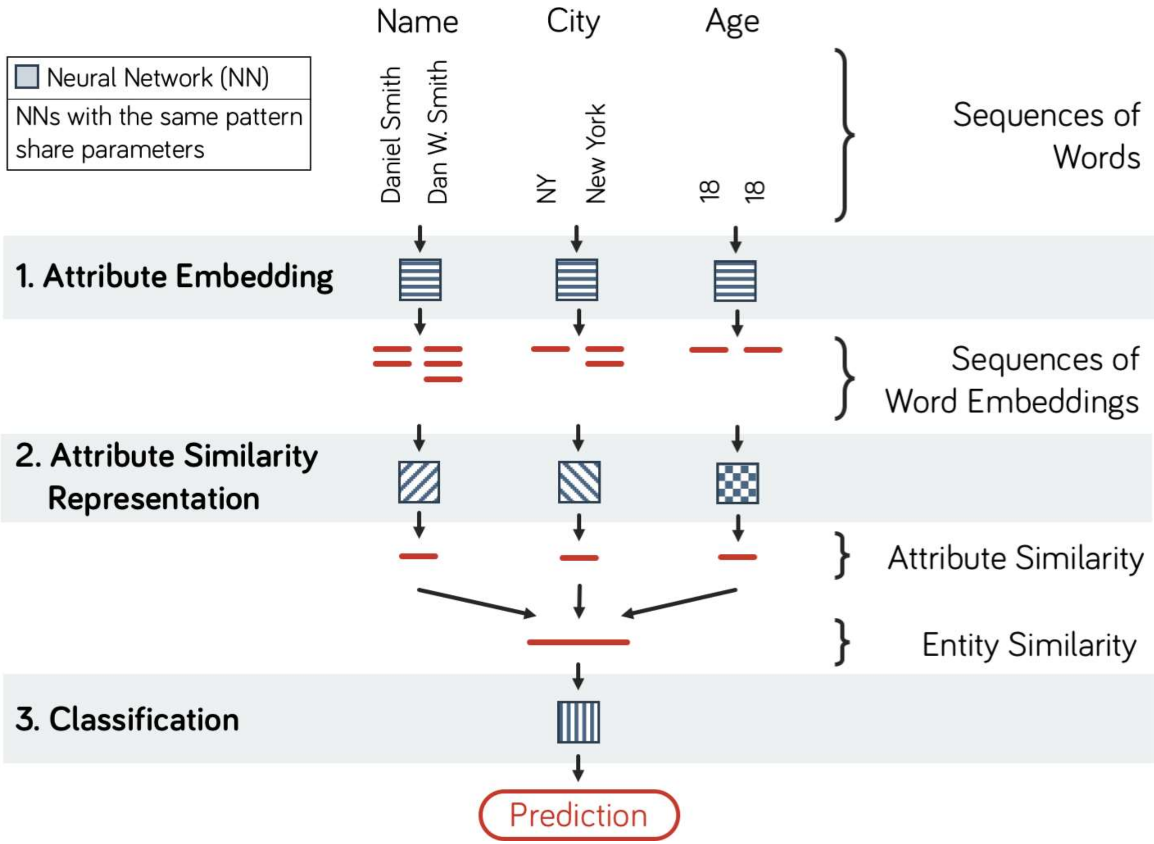
WDC Training Sets

Category	Size	Positives	Negatives
Computers	XLarge	9.690	58.771
	Large	6.146	27.213
	Medium	1.762	6.332
	small	722	2.112
Cameras	XLarge	7.178	35.099
	Large	3.843	16.193
	Medium	1.108	4.147
	Small	486	1.400

Category	Size	Positives	Negatives
Watches	XLarge	9.264	52.305
	Large	5.163	21.864
	Medium	1.418	4.995
	small	580	1.675
Shoes	XLarge	4.141	38.288
	Large	3.482	19.507
	Medium	1.214	4.591
	Small	530	1.533

http://webdatacommons.org/**largescaleproductcorpus**/v2/

DeepMatcher (2018)



- Embeddings: FastText
- Summarization: Bi-RNN with attention
- Similarity computation: element-wise difference and multiplication, concatenation
- Classification: Fully connected neural net, cross entropy loss

Mudgal, Sidharth, et al.: **Deep Learning for Entity Matching: A Design Space Exploration.** SIGMOD, 2018.

Evaluation: DeepMatcher versus Magellan

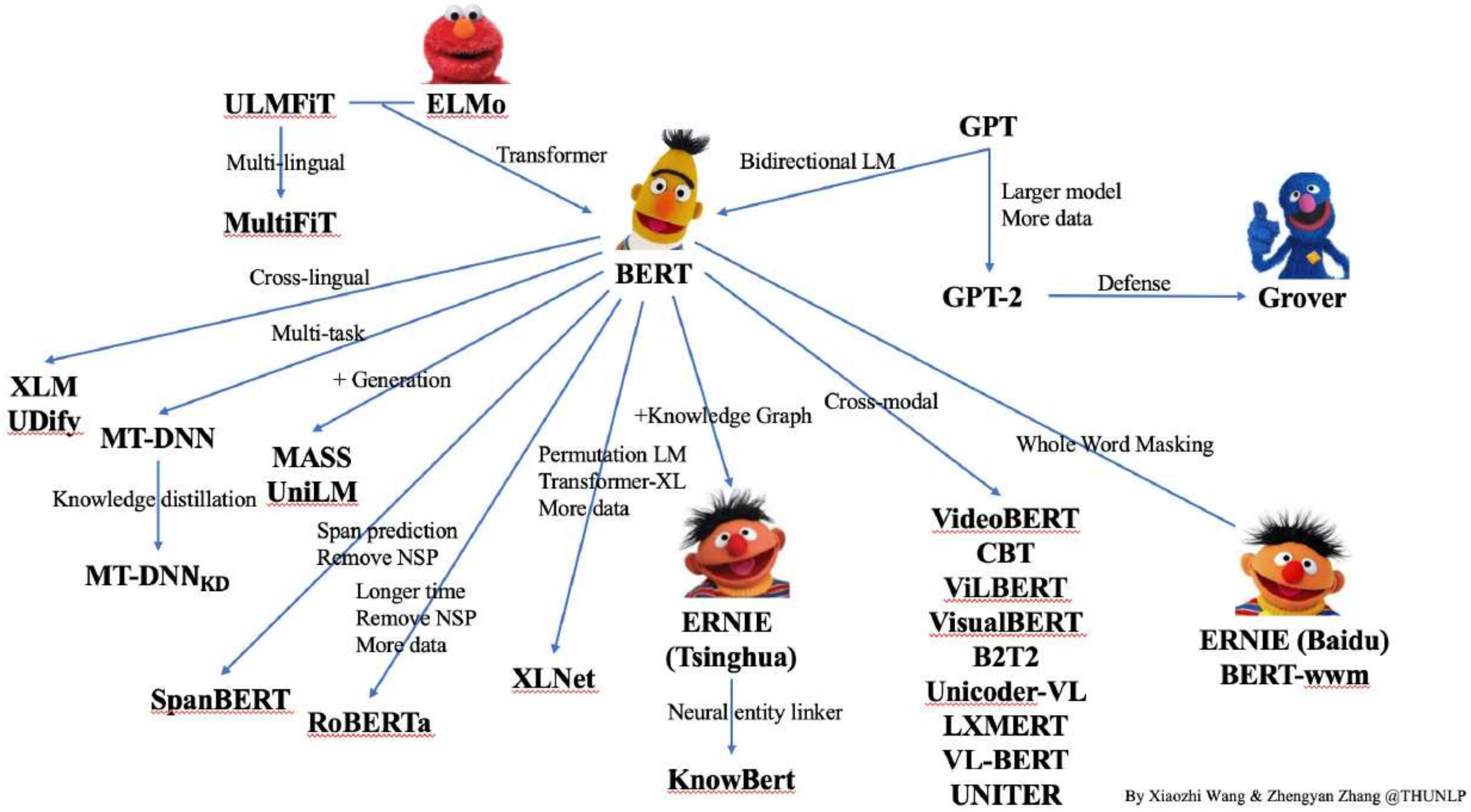
Type	Dataset	Magellan F1	DeepMatcher F1	Difference
Structured	iTunes-Amazon	91.2	88.5	-2.7
	DBLP-ACM	98.4	98.4	+0.0
	DBLP-Scholar	94.7	92.3	+2.4
	Walmart-Amazon	71.9	66.9	-5.0
Textual	Abt-Buy	43.6	62.8	+19.2
	Amazon-Google	49.1	69.3	+20.1
	WDC Computer - Large	64.5	89.5	+25.0
	WDC Computer - Small	57.6	70.5	+12.9

– DeepMatcher outperforms traditional methods on textual data
– mixed results on structured data

Konda, et al.: **Magellan: Toward Building Entity Matching Management**. PVLDB 2016.

Transformers started to win all Benchmarks in NLP

– Self-supervised pre-training on large text corpora
– Fine-tuning for downstream tasks



https://**huggingface.co**/docs/transformers/index

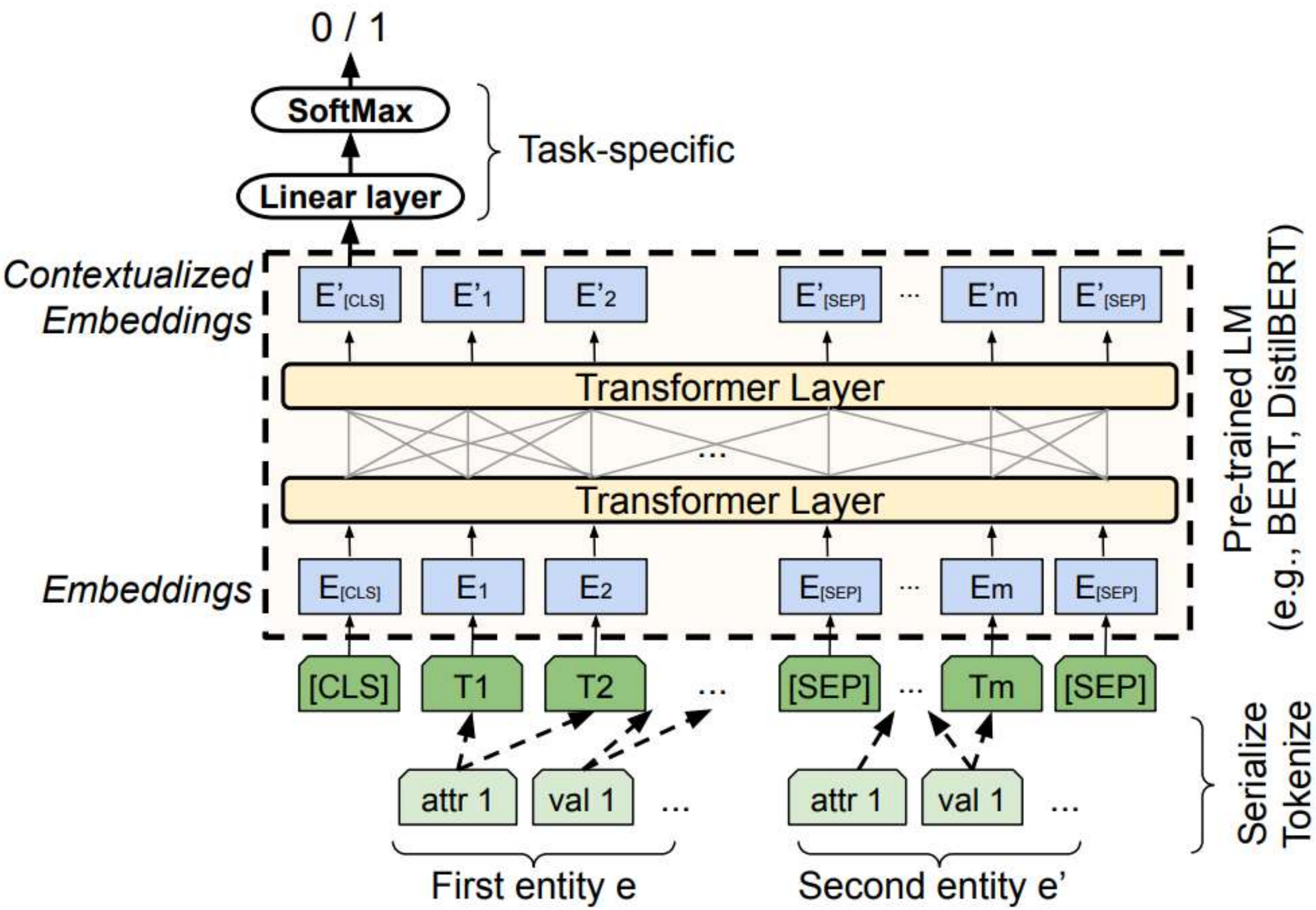
DITTO (2021)

- applies BERT, DistilBERT, RoBERTa for entity matching
- adds methods for entity summarization, highlighting matching clues, training data augmentation

- Entity serialization for BERT
  - Pair of entity descriptions are turned into single sequence
  - [CLS] Entity Description 1 [SEP] Entity Description 2 [SEP]
  - Entity Description = [COL] $attr_1$ [VAL] val_1 . . . [COL] $attr_k$ [VAL] val_k

Yuliang, et al: **Deep entity matching with pre-trained language models.** PVLDB 2021.

DITTO: Architecture



– [CLS] token summarizes the pair of entities
– linear layer on top of [CLS] token for matching decision

DITTO: Evaluation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Type	Dataset	DITTO F1	DeepMatcher F1	Magellan F1
Structured	iTunes-Amazon	97.0	88.5 +8.5	91.2 +5.8
	DBLP-ACM	99.0	98.4 +0.6	98.4 +0.6
	DBLP-Scholar	95.6	92.3 +3.3	94.7 +0.9
	Walmart-Amazon	86.8	66.9 +19.9	71.9 +14.9
Textual	Abt-Buy	89.3	62.8 +26.5	43.6 +45.7
	Amazon-Google	75.6	69.3 +6.3	49.1 +26.5
	WDC Computer - Large	91.7	89.5 +3.2	64.5 +27.2
	WDC Computer - Small	80.8	70.5 +10.3	57.6 +23.2

- constant improvement for structured data
- large performance gain for textual data

Potential Reasons for the Performance Gain

- Serialization allows to pay attention to all attributes
  - no strict separation between attributes
- WordPiece tokenizer breaks unknown terms into pieces
  - no problems with out of vocabulary terms
- Transfer learning from pre-training texts
  - different surface forms are already close in embedding space
- Contextualization of the embeddings
  - potentially more suited for capturing differing semantics

DITTO: Evaluation on WDC Datasets

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Size	xLarge (1/1)		Large (1/2)		Medium (1/8)		Small (1/20)	
Methods	DM	Ditto	DM	Ditto	DM	Ditto	DM	Ditto
Computers	90.80	95.45	89.55	91.70	77.82	88.62	70.55	80.76
	+4.65		+2.15		+10.80		+10.21	
Cameras	89.21	93.78	87.19	91.23	76.53	88.09	68.59	80.89
	+4.57		+4.04		+11.56		+12.30	
Watches	93.45	96.53	91.28	95.69	79.31	91.12	66.32	85.12
	+3.08		+4.41		+11.81		+18.80	
Shoes	92.61	90.11	90.39	88.07	79.48	82.66	73.86	75.89
	-2.50		-2.32		+3.18		+2.03	
All	90.16	94.08	89.24	93.05	79.94	88.61	76.34	84.36
	+3.92		+3.81		+8.67		+8.02	

- using BERT results in larger gains for small fine-tuning sets
- DITTO is more training efficient than DeepMatcher

Impact of the Base Transformer

- BERT: Pretrained on Books Corpus, Wikipedia
- RoBERTa: Pretrained on Books Corpus, Wikipedia, Common Crawl

Testset	Training Set	BERT	RoBERTa
WDC computers	xlarge	94.57	94.73
	large	92.11	94.68
	medium	89.31	**91.90**
	small	80.46	**86.37**
WDC cameras	xlarge	91.42	94.39
	large	91.02	93.91
	medium	87.02	**90.20**
	small	77.47	**85.74**
WDC watches	xlarge	95.76	94.87
	large	95.23	93.93
	medium	89.00	**92.28**
	small	78.73	**87.16**
abt-buy	default	84.64	**91.05**
dblp-scholar	default	95.27	**95.29**
company	default	91.70	91.81

- RoBERTa performs better for product matching
- base model matters for smaller fine-tuning sets

Seen versus Unseen during Training

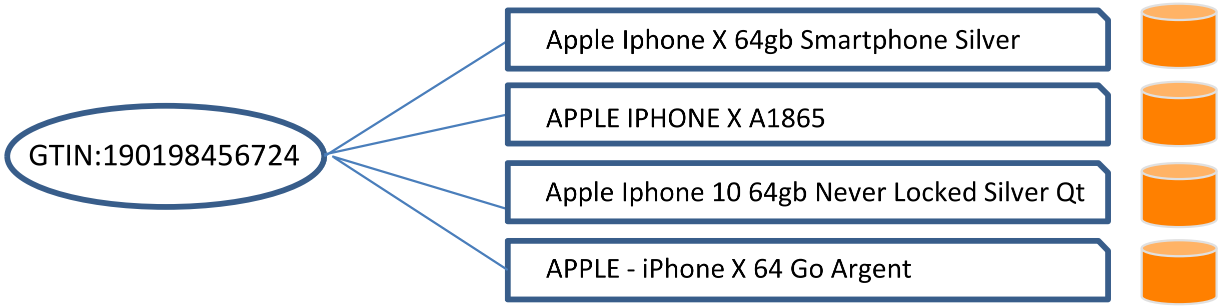
1. model trained on WDC computers large
2. model tested using different test sets:

Test Set	RoBERTa F1	BERT F1	DeepMatcher F1	Magellan F1
a) Seen products - different offers	94.73	92.11	89.55	63.45
b) Unseen products - high similarity	83.18	84.53	58.64	27.89
c) Unseen products - low similarity	77.72	76.92	58.78	59.04
Δ between a) and c)	-17.01	-15.19	-30.91	-4.41

Peeters, Bizer: **Dual-Objective Fine-Tuning of BERT for Entity Matching**. PVLDB 2021

Grouping Entity Descriptions

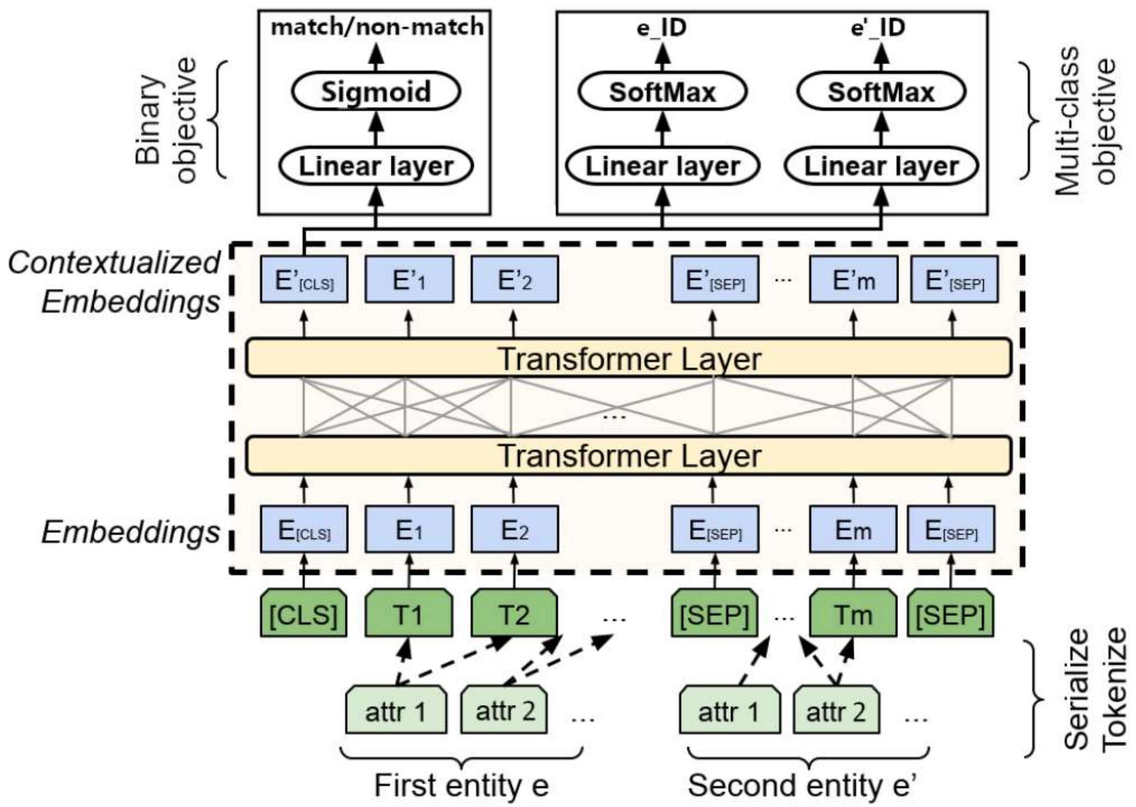
- **Shared identifiers** allow grouping of entity descriptions from different sources



- **Groups of entity descriptions** allow to
  - learn recognizers for single entities
  - treat entity matching as a multi-class classification task

JointBERT (2021)

– hypothesis: Learning to recognize single entities might also help to classify entity pairs
– combines pairwise matching and multi-class classification via multi-task learning



Overall loss:

$$L_i = BCEL\left(y_{b_i}, \hat{y}_{b_i}\right) + \left(CEL\left(y_{l_i}, \hat{y}_{l_i}\right) + CEL\left(y_{r_i}, \hat{y}_{r_i}\right)\right)$$

BCEL: Binary cross entropy loss
CEL: Cross entropy loss

Peeters, Bizer: **Dual-Objective Fine-Tuning of BERT for Entity Matching**. PVLDB 2021.

JointBERT: Evaluation

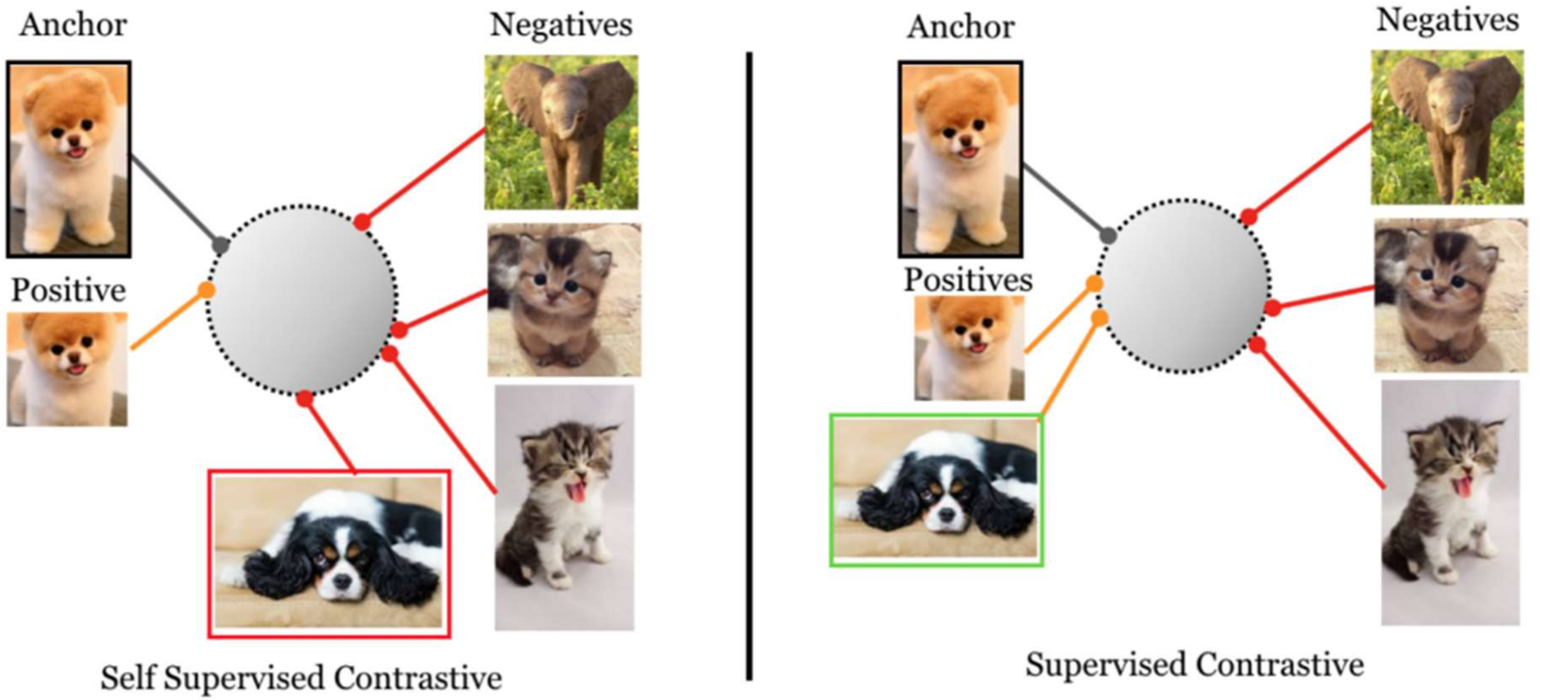
UNIVERSITY OF MANNHEIM
Data and Web Science Group

Testset	Training Set	Word Co-oc	Magellan	Deepmatcher	BERT	RoBERTa	Ditto	JointBERT
WDC computers	xlarge	82.39	63.16	88.95	94.57	94.73	96.53	**97.49**
	large	81.23	64.56	84.32	92.11	94.68	93.81	**96.90**
	medium	70.94	61.59	69.85	89.31	**91.90**	88.97	88.82
	small	62.69	57.60	61.22	80.46	**86.37**	81.52	77.55
WDC cameras	xlarge	73.33	51.70	84.88	91.42	94.39	94.74	**98.02**
	large	76.24	54.49	82.16	91.02	93.91	94.41	**96.51**
	medium	69.89	54.99	69.34	87.02	**90.20**	87.97	87.91
	small	64.86	52.78	59.65	77.47	**85.74**	78.67	78.30
WDC watches	xlarge	79.78	56.04	88.34	95.76	94.87	97.05	**97.09**
	large	79.64	60.59	86.03	95.23	93.93	97.17	**98.46**
	medium	69.54	66.62	67.92	89.00	**92.28**	89.16	87.46
	small	63.49	59.73	54.97	78.73	**87.16**	81.32	75.83
WDC shoes	xlarge	70.38	61.45	86.74	87.44	88.88	93.28	**97.88**
	large	71.18	60.48	83.17	87.37	86.60	90.07	**95.16**
	medium	72.43	59.80	74.40	79.82	81.12	**83.20**	82.61
	small	63.65	58.57	64.71	74.49	**80.29**	75.13	73.13
DI2KG monitor-seen	default	83.59	21.79	77.41	96.22	96.65	86.51	**97.82**
DI2KG monitor-unseen	default	21.24	16.02	33.93	91.49	**93.26**	73.52	82.86
DI2KG monitor-combined	default	33.82	16.90	37.37	92.19	**93.76**	75.28	84.77
abt-buy	default	36.30	43.60	62.80	84.64	**91.05**	82.11	83.44
dblp-scholar	default	85.38	92.30	94.70	95.27	**95.29**	94.47	93.99
company	default	71.81	79.80	**92.70**	91.70	91.81	90.68	91.40

– increase of 1 to 5% F1 given enough training examples

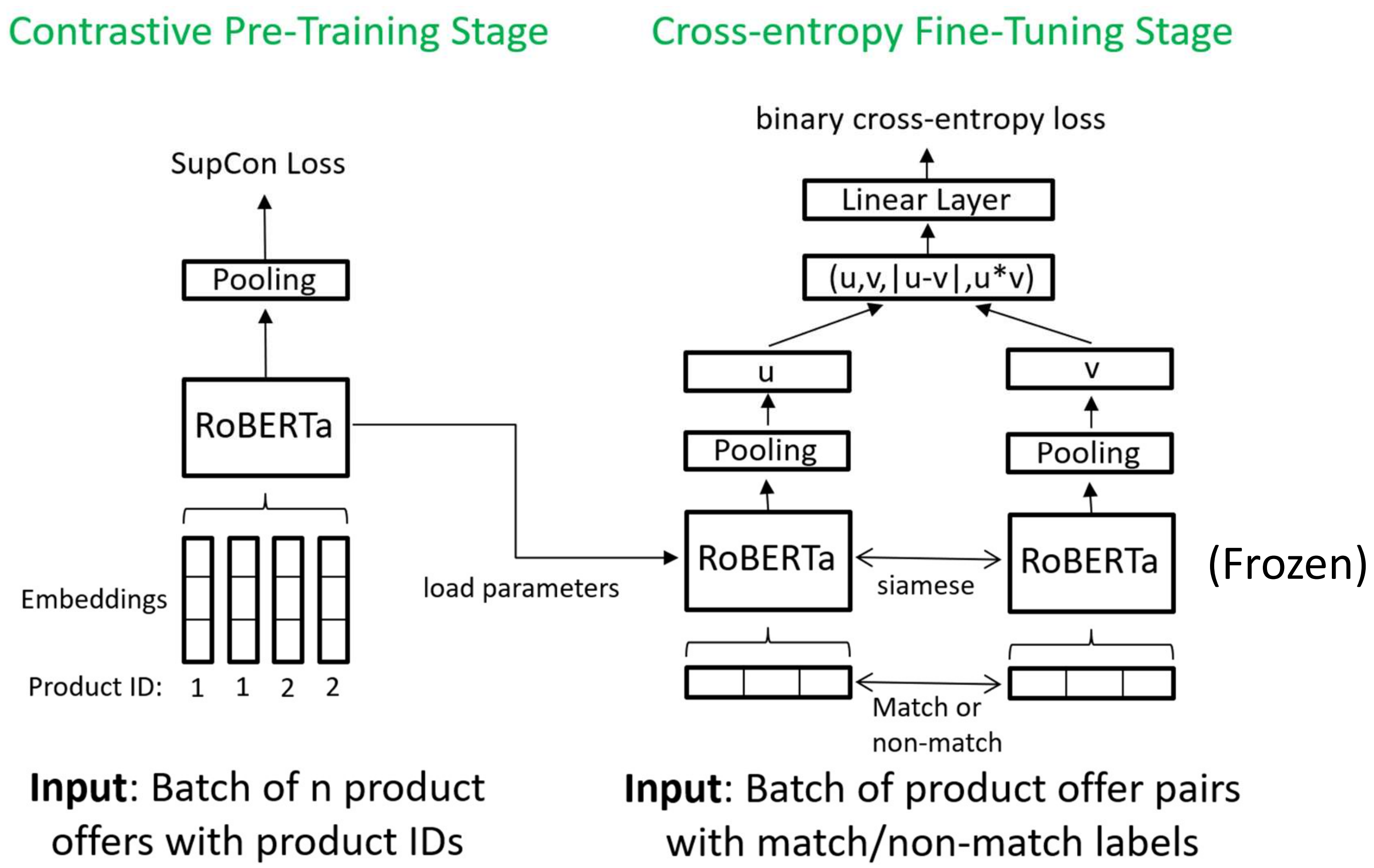
Contrastive Pretraining in Vision

- maximizes distance between classes in the embedding space
- uses large batches containing many positive and negative examples



Khosla, et al.: **Supervised Contrastive Learning**. NeurIPS 2020.

Supervised Contrastive Pretraining for Entity Matching (2022)



Peeters, Bizer: **Supervised Contrastive Learning for Product Matching**. WWW Companion 2022.

Evaluation: Supervised Contrastive Pretraining R-SupCon

UNIVERSITY OF MANNHEIM
Data and Web Science Group

	Abt-Buy	Amazon-Google	WDC Computers			
# Training Pairs	~7.5K	~9K	~3K (small)	~8K (medium)	~23K (large)	~68K (xlarge)
DeepMatcher	62.80	70.70	61.22	69.85	84.32	88.95
RoBERTa	91.05	74.10	86.37	91.90	94.68	94.73
Ditto	89.33	75.58	80.76	88.62	91.70	95.45
JointBERT	-	-	77.55	88.82	96.90	97.49
R-SupCon	93.70	**79.28**	93.18	97.66	98.16	**98.33**
R-SupCon+augmen	**94.29**	76.14	**95.21**	**98.50**	**98.50**	98.33
Δ to best baseline	+ 3.24	+ 3.70	+ 8.84	+ 6.60	+ 1.60	+ 0.84

Improvement also for small training sets in contrast to JointBERT

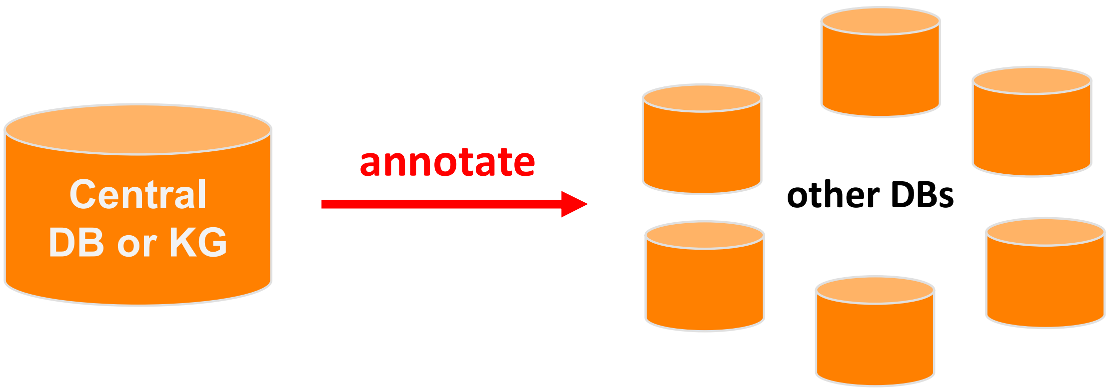
Summary: Deep Learning for Entity Matching

- Transformer-based matchers boost performance
  - huge increase for textual data
  - significant increases for structured data
- reduced feature engineering effort
  - less value normalization necessary due to pre-training
  - less information extraction effort due to serialization
- requirements for training data
  - pre-training data should be close to the downstream use case
  - all entities should be covered with training examples
  - medium size training sets are enough for fine-tuning

2. Schema Matching

Schema Matching

- Deep learning-based methods treat schema matching as
  1. column type annotation
  2. relation extraction
- in a one-to-n scenario
  - central DB or KG defines column types and relations
  - tables in other DBs are annotated with these types and relations

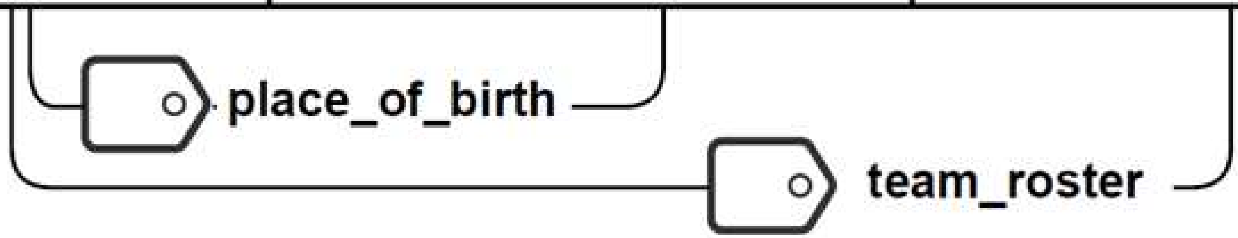


Column Type Annotation

film	director	producer	country
???	???	???	???
Happy Feet	George Miller, Warren Coleman, Judy Morris	Bill Miller, George Miller, Doug Mitchell	USA
Cars	John Lasseter, Joe Ranft	Darla K. Anderson	UK
Flushed Away	David Bowers, Sam Fell	Dick Clement, Ian La Frenais, Simon Nye	France

Relation Extraction

person	location	sports_team
Max Browne	Sammamish, Washington	Southern California
Thomas Tyner	Aloha, Oregon	Oregon
Derrick Henry	Yulee, Florida	Alabama



Evaluation Dataset: Wikitables

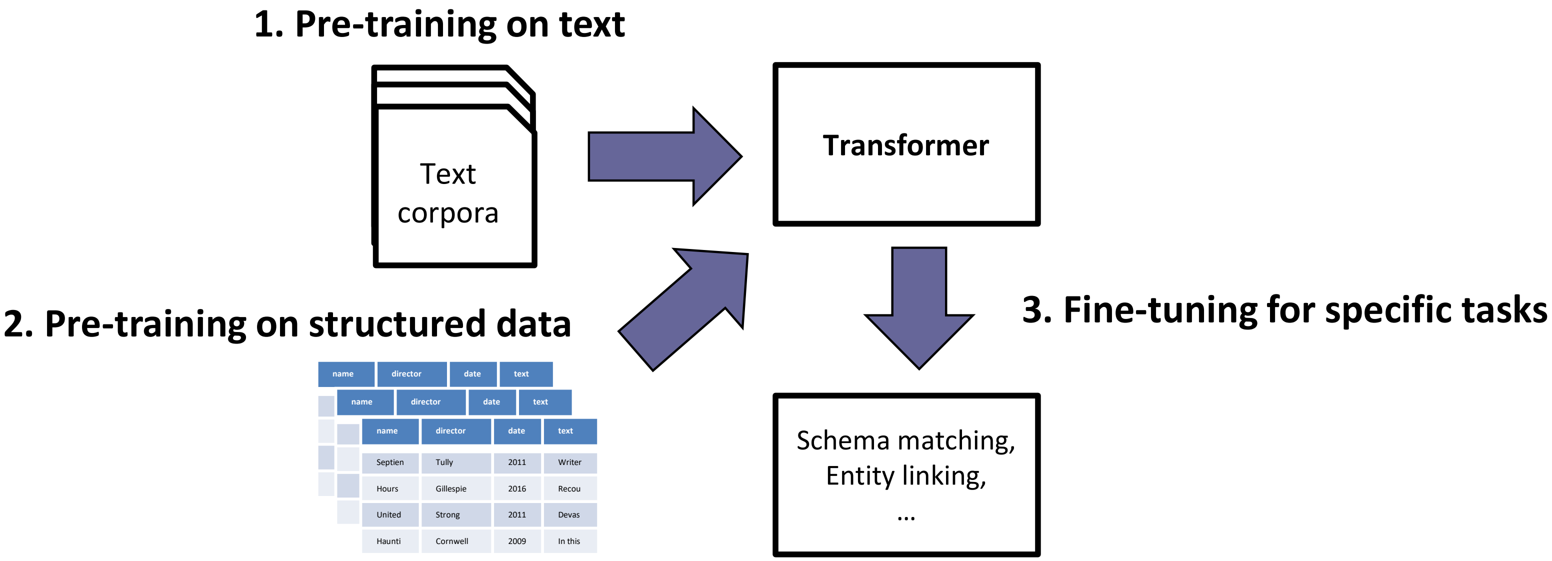
- Corpus of 570.000 tables from Wikipedia
  - containing links to Wikipedia pages (entity identification)

- Column Type Annotation Task
  - 255 type from Freebase assigned by exploiting entity links in tables
  - 400.000 tables train / 5000 dev / 5000 test
  - task modeled as binary multi-label task → each column 255 times yes/no
- Relation Extraction Task
  - 121 relations from Freebase assigned by exploiting entity links in tables
  - 50.000 tables train / 1500 dev / 1500 test
  - task modeled as binary multi-label task → each column 121 times yes/no

Deng, et al.: **TURL: Table Understanding through Representation Learning**. PVLDB 2020.

TURL (2020)

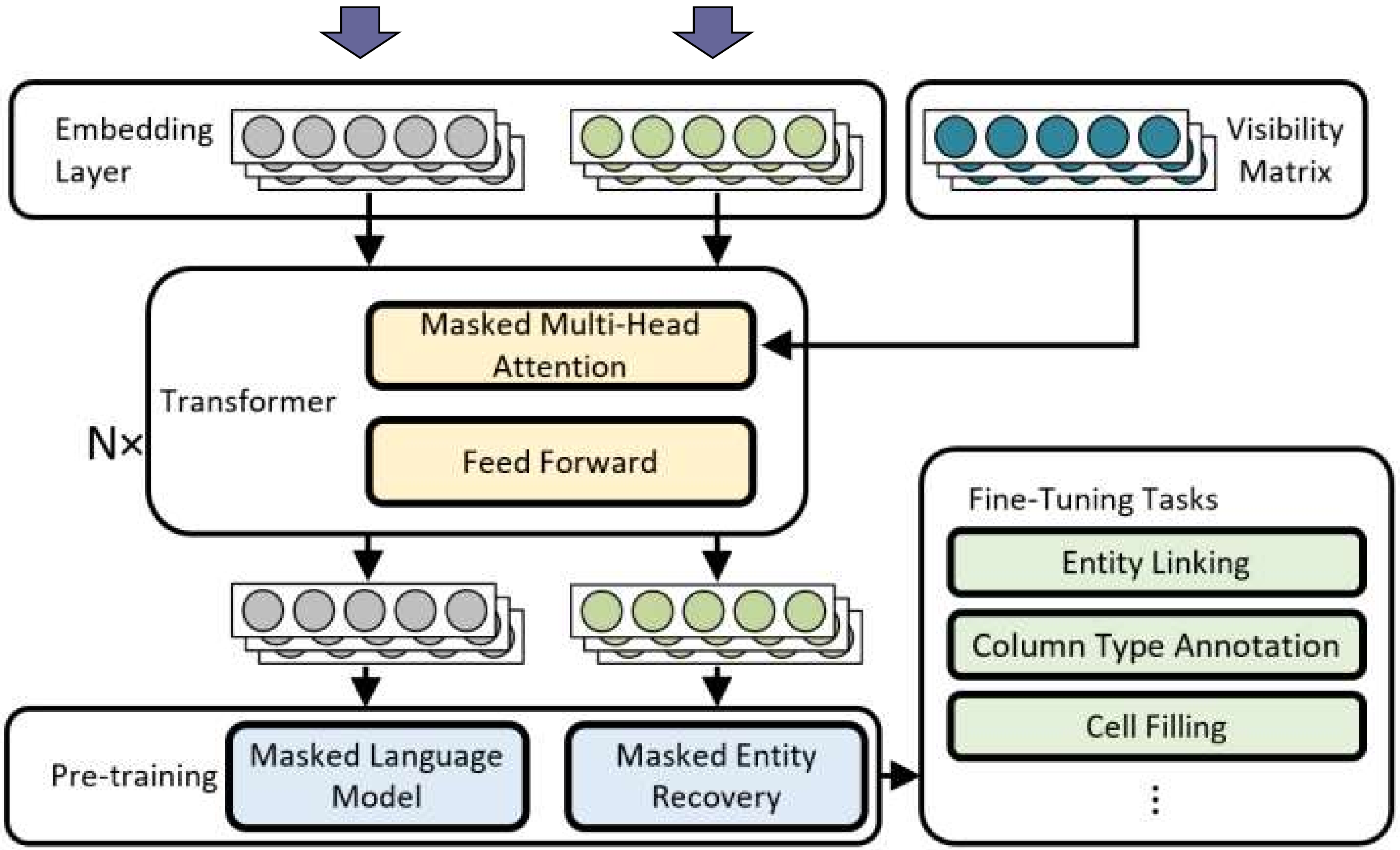
- aims at learning generic table representations that are useful across a wide range of tasks
  - Pre-training: Self-supervised table representation learning
  - Fine-tuning: For 6 specific downstream tasks



Deng, et al.: **TURL: Table Understanding through Representation Learning**. PVLDB 2020.

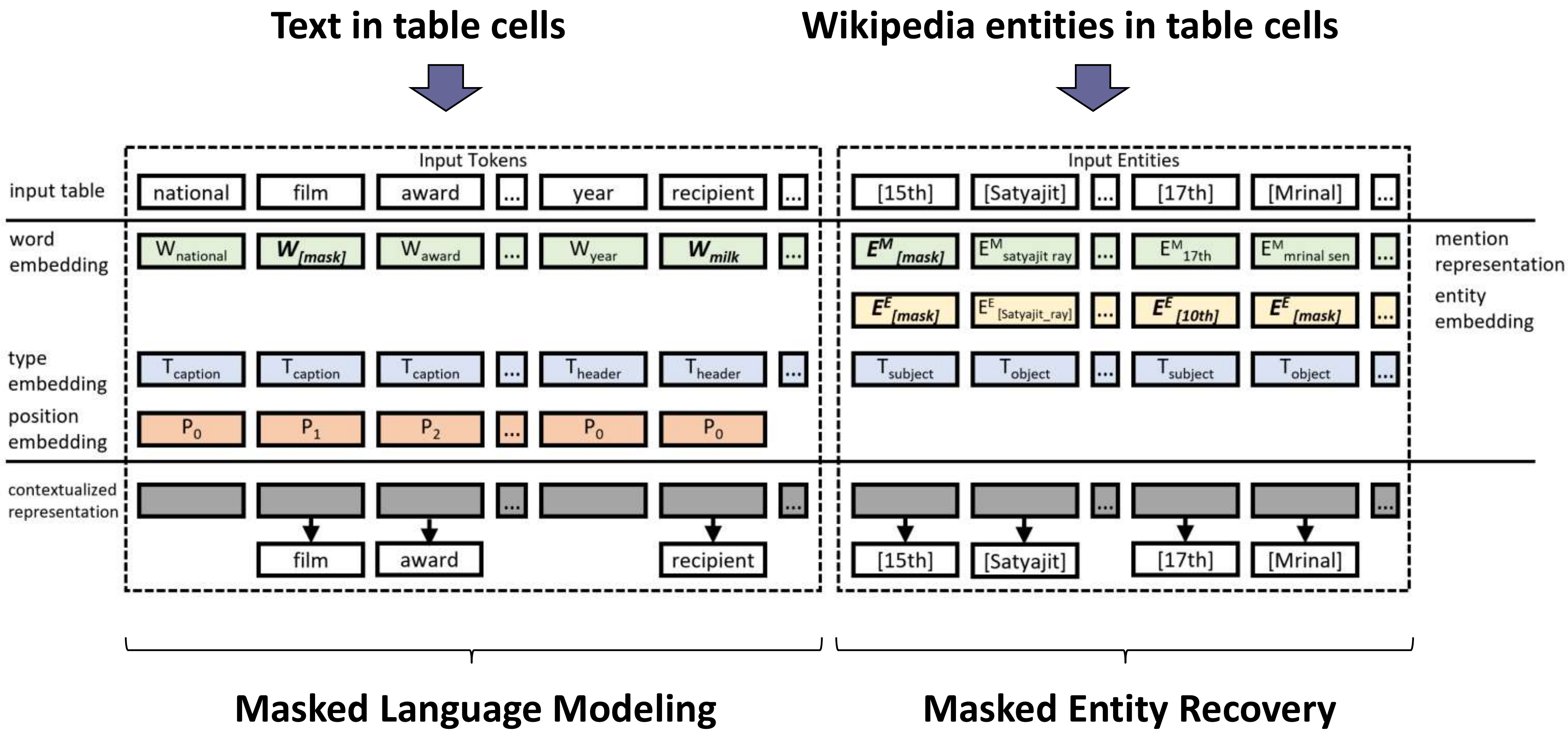
TURL: Overview

Text in table cells **Wikipedia entities in table cells**



TURL: Table Embedding and Pre-Training

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Pooling of cell text into single embedding allow to fit many cells into 512 token limit

TURL: Attention Masking

- Visibility matrix to impose table structure in the encoder layers
- Highlighted token can only pay attention to visible tokens/entities

(a) Visibility for *caption tokens*

name	director	date	description

(b) Visibility for header tokens

name	director	date	description
	Michael Tully		
	Craig Gillespie		
	James Strong		
	Peter Cornwell		

(c) Visibility for cell entities

	director		
	Michael Tully		
	Craig Gillespie		
United	James Strong	2011	A devastating …
	Peter Cornwell		

TURL: Evaluation

– Column Type Annotation

Method	F1	P	R
Sherlock (only entity mentions)	78.47	88.40	70.55
TURL + fine-tuning (only entity mentions)	88.86	90.54	87.23
TURL + fine-tuning	94.75	94.95	94.56
Largest Δ to Sherlock	16.28	6.55	24.01

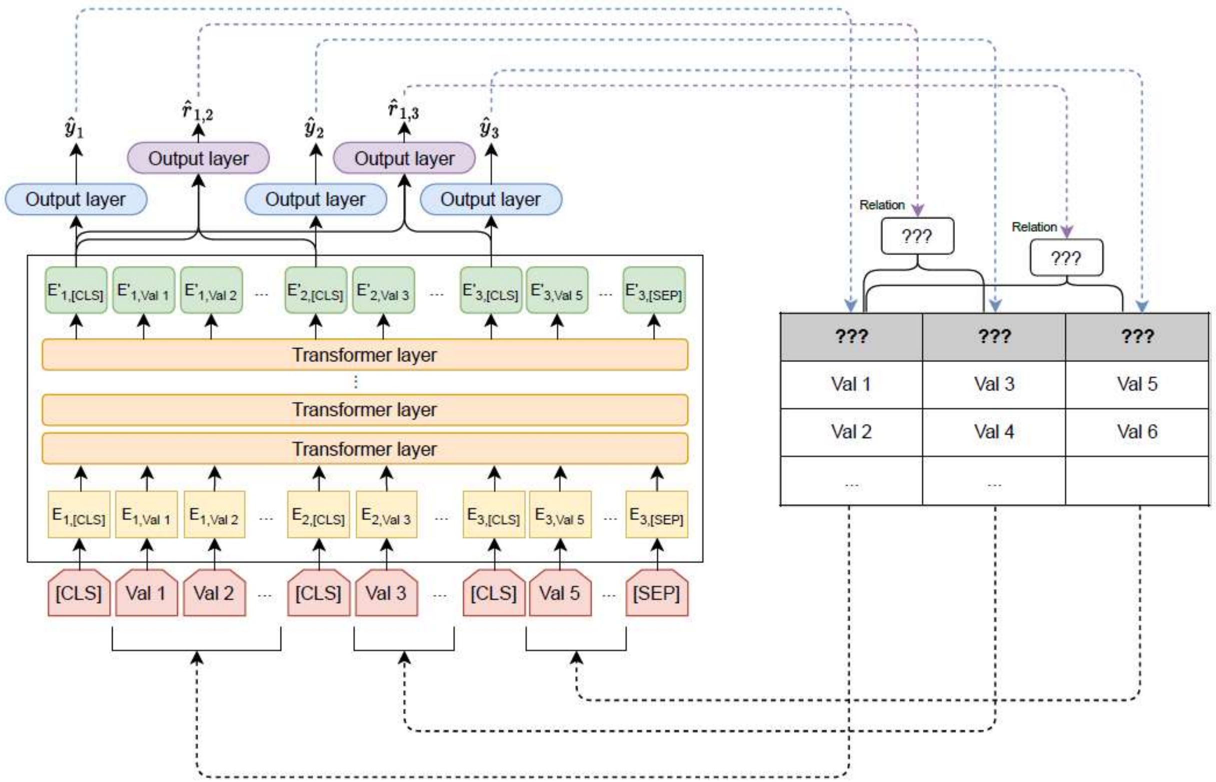
– Relation Extraction

Method	F1	P	R
BERT	90.94	91.18	90.69
TURL + fine-tuning (only table metadata)	92.13	91.17	93.12
TURL + fine-tuning	94.91	94.57	95.25
Largest Δ to Sherlock	+3.97	+3.39	+4.29

Hulsebos, et al.: **Sherlock: A Deep Learning Approach to Semantic Data Type Detection**. KDD 2019.

DoDuo (2022)

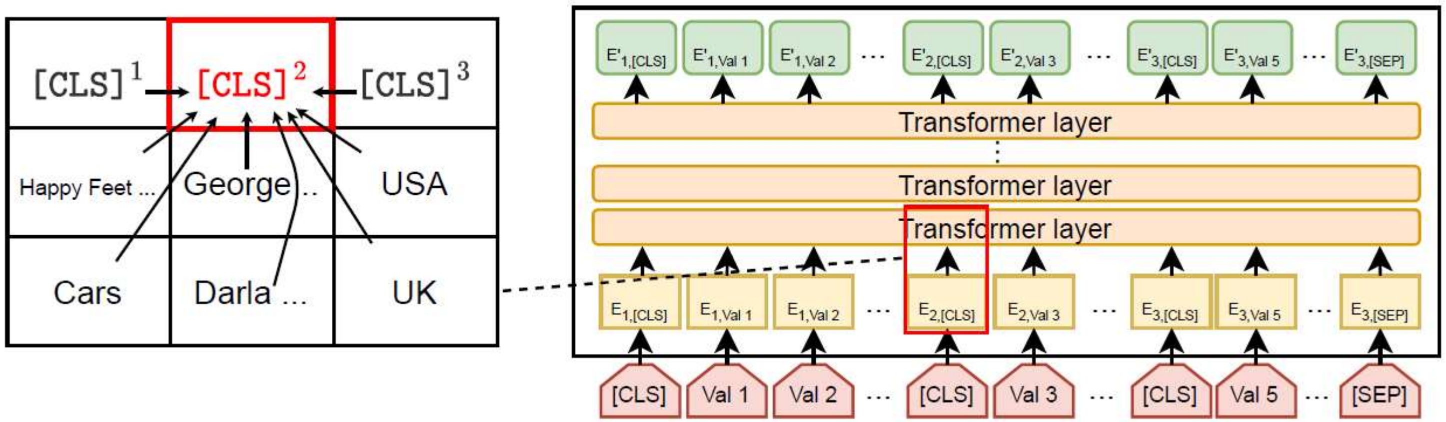
– directly fine-tunes BERT for column and relation annotation tasks
– exploits synergies between tasks using multi-task learning



Suhara, et al.: **Annotating Columns with Pre-trained Language Models**. SIGMOD 2022.

DoDuo: Cell Attention

– a table cell can pay attention to all neighboring cells



– in contrast to visibility masking in TURL

[redacted]	**director**	[redacted]	[redacted]
[redacted]	Michael Tully	[redacted]	[redacted]
[redacted]	Craig Gillespie	[redacted]	[redacted]
United	James Strong	2011	A devastating …
[redacted]	Peter Cornwell	[redacted]	[redacted]

DoDuo: Evaluation

– Column Type Annotation

Method	F1	P	R
Sherlock	78.47	88.40	70.55
TURL (DistilBERT)	88.86	90.54	87.23
DoDuo (BERT)	92.45	92.45	92.21
Δ to TURL	+3.59	+2.09	+4.98

– Relation Extraction

Method	F1	P	R
TURL (DistilBERT)	90.94	91.18	90.69
DoDuo (BERT)	91.72	91.97	91.47
Δ to TURL	+0.78	+0.79	+0.88

Further Approaches

System	Base-Architecture	Evaluated Downstream Tasks	Pre-Training Objectives
TURL (Google)	TinyBERT	Entity linking, column typing, relation extraction , entity expansion, cell filling, schema augmentation	MLM / masked entity recovery
DoDuo (Megagon)	BERT	column typing, relation extraction	MLM
TUTA (Microsoft)	BERT	Cell and table type classification	MLM / cell-level cloze / table-context retrieval
TCN (Amazon)	CNN	Column typing, relation extraction	MLM
TaBERT (Facebook)	BERT	Semantic parsing	MLM / masked column prediction / cell value recovery
TAPAS (Google)	BERT	Semantic parsing	MLM
TABBIE (Sony)	BERT	cell type, row and column population	MLM / corrupt cell detection

Pujara, et al.: **From Tables to Knowledge: Recent Advances in Table Understanding**. Tutorial at KDD2021.

Summary:
Deep Learning for Schema Matching

- Significant performance gains
  - on both column type annotation and relation extraction tasks

- Table embedding is an active research area
  - generic vs. task-specific embeddings
  - a wide range of table related tasks are explored

- Small number of evaluation datasets
  - results might include dataset bias

Conclusions

1. Transformer-based methods **boost matching performance**
   – huge improvements for entity matching
   – significant gains for schema matching
2. Transformer-based methods **reduce feature engineering effort**
   – less value normalization effort due to pre-training
   – less information extraction effort due to entity serialization
3. **Domain-relatedness of training data** is key to success
   – potential of using the Web as a source of training data

Hands On +
Staying Up to Date

- The source code is available for all discussed papers
  - you can test if the methods work for your use cases
  - we are happy to help you

- Benchmark results are collected on Papers with Code
  - good reference point
    for latest developments

https://paperswithcode.com/task/entity-resolution/
https://paperswithcode.com/task/table-annotation/

Thank you.

eMail: christian.bizer@uni-mannheim.de

Web: https://www.uni-mannheim.de/dws/people/professors/prof-dr-christian-bizer/