RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation

Nan Tang
QCRI, HBKU, Qatar
ntang@hbku.edu.qa

Ju Fan
Renmin University, China
fanj@ruc.edu.cn

Fangyi Li
Renmin University, China
fangyili@ruc.edu.cn

Jianhong Tu
Renmin University, China
tujh@ruc.edu.cn

Xiaoyong Du
Renmin University, China
duyong@ruc.edu.cn

Guoliang Li
Tsinghua University, China
liguoliang@tsinghua.edu.cn

Sam Madden
CSAIL, MIT, USA
madden@csail.mit.edu

Mourad Ouzzani
QCRI, HBKU, Qatar
mouzzani@hbku.edu.qa

ABSTRACT

Can AI help automate human-easy but computer-hard data preparation tasks that burden data scientists, practitioners, and crowd workers? We answer this question by presenting RPT, a denoising autoencoder for *tuple-to-X* models ("*X*" could be tuple, token, label, JSON, and so on). RPT is pre-trained for a *tuple-to-tuple* model by corrupting the input tuple and then learning a model to reconstruct the original tuple. It adopts a Transformer-based neural translation architecture that consists of a bidirectional encoder (similar to BERT) and a left-to-right autoregressive decoder (similar to GPT), leading to a generalization of both BERT and GPT. The pre-trained RPT can already support several common data preparation tasks such as data cleaning, auto-completion and schema matching. Better still, RPT can be fine-tuned on a wide range of data preparation tasks, such as value normalization, data transformation, data annotation, etc. To complement RPT, we also discuss several appealing techniques such as collaborative training and few-shot learning for entity resolution, and few-shot learning and NLP question-answering for information extraction. In addition, we identify a series of research opportunities to advance the field of data preparation.

PVLDB Reference Format:
Nan Tang, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Sam Madden, and Mourad Ouzzani. RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation. PVLDB, 14(8): 1241 - 1248, 2021.
doi:10.14778/3457390.3457391

1 INTRODUCTION

Data preparation — including data cleaning [1], data transformation [31], entity resolution [24], information extraction [10], and so forth — is the most time-consuming and least enjoyable work for data scientists [18]. Next, we present several scenarios to better understand these problems.

Scenario 1: Data Cleaning. Figure 1(a) Q1 and Q2 show two typical data cleaning problems. (i) *Cell Filling:* Question Q1 asks for the **city** for the "Michael Jordan" whose expertise is "Machine Learning". (ii) *Value Filling:* Q2 asks for the last name of someone who works at "CSAIL MIT" with the first name "Michael".

Q1: *r1*[name, expertise, city] = (Michael Jordan, Machine Learning, **[M]**)
A1: Berkeley

Q2: *r3*[name, affiliation] = (Michael **[M]**, CSAIL MIT)
A2: Cafarella

Q3: *r2*[name, expertise, **[M]**] = (Michael Jordan, Basketball, New York City)
A3: city

(a) Sample Tasks for Value Filling ([M]: value to fill)

	product	company	year	memory	screen
e1	iPhone 10	Apple	2017	64GB	5.8 inchs
e2	iPhone X	Apple Inc	2017	256GB	5.8-inch
e3	iPhone 11	AAPL	2019	128GB	6.1 inches

(b) A Sample Entity Resolution Task

	type	description	label
s1	notebook	2.3GHz 8-Core, 1TB Storage, 8GB memory, 16-inch Retina display	8GB
t1	phone	6.10-inch touchscreen, a resolution of 828x1792 pixels, A14 Bionic processor, and come with 4GB of RAM	4GB

(c) A Sample Information Extraction Task (*s1*: example, *t1*: task)

Figure 1: Motivating Scenarios.

Answering Q1 can help solve a series of problems such as error detection, data repairing, and missing value imputation; and answering Q2 can help auto-completion (*e.g.,* give the answer A2 "Cafarella") and auto-suggestion (*e.g.,* provide a list of candidate names such as {Cafarella, Stonebraker}).

Scenario 2: Attribute Filling for Schema Matching. Figure 1(a) Q3 asks for the attribute name for the value "New York City", *w.r.t.* **name** "Michael Jordan" and **expertise** "Basketball". Answering this question can help schema matching, a core data integration problem [21], by better aligning attributes from different tables.

Scenario 3: Entity Resolution (ER) Figure 1(b) shows a typical ER task that asks whether e_1, e_2 and e_3 are the "same".

A human with enough knowledge can tell that "iPhone 10" = "iPhone X" ≠ "iPhone 11", "Apple" = "Apple Inc" = "AAPL", and "inches" = "-inch". Hence, one can decide that e_1 and e_2 do not match e_3, and e_1 matches e_2 (if the memory does not matter).

Scenario 4: Information Extraction (IE) Figure 1(c) shows an IE task, which is typically done via crowdsourcing [39]. A requester

* Ju Fan is the corresponding author.

Proceedings of the VLDB Endowment, Vol. 14, No. 8 ISSN 2150-8097.
doi:10.14778/3457390.3457391

provides several samples (*e.g.,* s_1) that show what "label" should be extracted, and asks workers to perform similar tasks (*e.g.,* t_1).

A crowd worker needs first to interpret the task by analyzing s_1 (and maybe a few more examples) and concretizes it as "*what is the memory size*". Afterwards, he can perform t_1 by extracting the label 4GB from t_1[`description`] by knowing "RAM" is for memory.

Challenges. Scenarios (1–4) are simple for humans, but are hard for computers. To solve them, computers face the following challenges. (1) *Knowledge:* computers need to have the background knowledge through *understanding* an enormous corpora of tables. (2) *Experience:* computers should be able to learn from prior and various tasks. (3) *Adaptation:* computers should (quickly) adjust to new inputs and new tasks.

Vision. Indeed, these problems have been seen as 'holy grail' problems for the database community for decades [1, 21, 25, 30, 36, 46], but despite thousands of papers on these topics, they still remain unsolved. Recent evidence from the NLP community, where DL-based models and representations have been shown to perform nearly as well as humans on various language understanding and question answering tasks, suggests that a learned approach may be a viable option for these data preparation tasks as well.

Desiderata. The desiderata for AI-powered tools to achieve near-human intelligence for data preparation is summarized as below, in response to Challenges (1–3), respectively. (1) *Deep learning architecture and self-supervised pre-training*. We need a deep learning architecture that can learn from many tables, similar to language models that can learn from a large text corpora. This simulates how humans gain knowledge. Moreover, it should be pre-trained without human provided labels, *i.e.,* self-supervision. (2) *Transfer learning*. The model should be able obtain knowledge from different tasks on different datasets. (3) *Fine-tuning and few-shot learning*. The pre-trained model should allow customization on different downstream applications through fine-tuning. In addition, it should be able to understand a new task from a few examples.

RPT Is *Almost* All You Need. The design of ***R***elational ***P***re-trained ***T***ransformer (RPT) is inspired by recent successes of DL models in NLP. The fundamental questions for relational data understanding on data preparation are: (1) *what is the architecture?* and (2) *what is the surrogate task for pre-training?*

(1) RPT Architecture. Typical choices are encoder-only such as BERT [20], decoder-only such as GPT-3 [8], or encoder-decoder such as BART [44] and T5 [55]. In fact, the encoder-decoder architecture can be considered as a generalization of the encoder-only model (*e.g.,* BERT) and the decoder-only model (*e.g.,* GPT-3). Recent studies from BART and T5 found that encoder-decoder models generally outperform encoder-only or decoder-only language models. Thus, a Transformer-based [64] encoder-decoder model provides more flexibility and can be adapted to a wide range of data preparation tasks, and hence can be used by RPT.

(2) RPT Pre-training. There have been several works on pre-training using tables, such as TAPAS [35], TURL [19], TaBERT [72] and TabFact [13]. However, since most data preparation tasks are in the granularity of tuples, instead of entire tables, we posit that training RPT tuple-by-tuple is more desirable. For the pre-training objectives, most recent studies confirm that *fill-in-the-blank* style denoising objectives (where the model is trained to recover missing pieces in the input) work best; examples include BERT [20], BART [44], and T5 [55] for NLP, and TURL [19] for relational tables.

Contributions. We make the following notable contributions.

- *RPT*: We describe a standard Transformer-based denoising autoencoder architecture to pre-train sequence-to-sequence models for *tuple-to-tuple* training, with new *tuple-aware masking* mechanisms. (Section 2)
- *Fine-tuning RPT:* We discuss a wide range of data preparation tasks that can be supported by fine-tuning RPT. (Section 3)
- *Beyond RPT:* We discuss several appealing techniques that can complement RPT in specific data preparation tasks, *e.g.,* collaborative training and few-shot learning for ER, and few-shot learning and NLP question-answering for IE. (Section 4)

2 RPT

2.1 Architecture

RPT uses a standard sequence-to-sequence (or encoder-decoder) Transformer [64] model, similar to BART [52], as shown in Figure 2.

Encoder. RPT uses a bidirectional encoder (similar to BERT [20]) because it has the advantage of learning to predict the corrupted data bidirectionally, from both the context on the left and the context on the right of the corrupted data. Moreover, it is Transformer-based, which can use self-attention to generate a richer representation of each input token. Hence, a Transformer-based bidirectional encoder is a natural fit for reading tuples where, by definition, the ordering of (attribute name, attribute value) pairs is irrelevant.

Decoder. RPT uses a left-to-right autoregressive decoder (similar to GPT-3 [8]).

2.2 Pre-training RPT

RPT is pre-trained on tuples, for which we just need to corrupt tuples and then optimize a reconstruction loss – the cross-entropy between the model output and the original tuple.

Tuple Tokenization. We represent each tuple as a concatenation of its attribute names and values. For example, tuple t_1 in Figure 1(a) can be tokenized as:

name Michael Jordan expertise Machine Learning city Berkeley

Token Embeddings. Because there is a clear semantic difference between attribute names and values, we can add special tokens, [**A**] before an attribute name and [**V**] before an attribute value. Token embeddings are widely used in NLP tasks, such as the [CLS] (indicating the start) and [SEP] (indicating the next sentence) tokens used by BERT [20]. Hence, we can get a sequence of t_1 with a richer tuple-aware semantics as:

[**A**] name [**V**] Michael Jordan [**A**] expertise [**V**] Machine Learning [**A**] city [**V**] Berkeley

Positional and Column Embeddings. We can add additional meta-data such as positional embeddings (*i.e.,* indicating the token's position in the sequence) and segment embeddings (*e.g.,* adding the same segment embedding to the tokens belonging to the same attribute value), which are inspired by TAPAS [35] for table parsing.

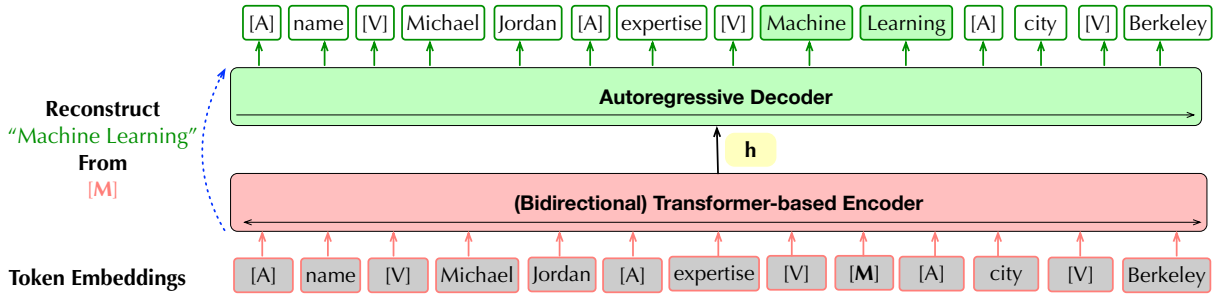

Figure 2: The RPT Architecture.

Working Mechanism. Given an input sequence of a tuple with some value to be masked out (*e.g.,* "Machine Learning" in Figure 2) represented by a "mask token" [**M**], along with rich semantic information (*e.g.,* an attribute [**A**] or a value [**V**], which position, and which column), the bidirectional encoder will look at the information before and after the masked token [**M**], learn which token to pay attention to (using Transformer [64]), and generate an intermediate vector representation h. The autoregressive decoder will take h as input, and generate an output sequence, by denoising the masked input [**M**] to be "Machine Learning". By doing so, we can train RPT in an unsupervised fashion, without any human labels.

One difficulty is to predict how many tokens are masked by one [**M**]. Note that, BERT [20] masks each token with one [**M**]; i.e., "Machine Learning" will be masked as [**M**][**M**], which tells explicitly how many tokens are missing. We cannot do the same (masking each token with one [**M**]), because during prediction, we do not know how many tokens are missing. The ability to teach the model to predict the length of the missing tokens masked by one single mask token [**M**] can be achieved by *text infilling* [40], which will be discussed next.

Token Masking. (1) *Attribute Name Masking:* We randomly select attribute names to mask, *e.g.,* **`name`**.

(2) *Entire Attribute Value Masking:* We randomly select entire attribute values to mask, *e.g.,* "Machine Learning" is masked with one [**M**] (see Figure 2), which forces RPT to first predict the number of masked tokens and then predict the tokens.

(3) *Single Attribute Value Masking:* We randomly select a single attribute value (*i.e.,* one token) to mask, *e.g.,* "Jordan".

Note that one possible optimization to the above process is as follows. Instead of giving RPT the full freedom to learn how input tokens attend on each other in the form of an attention matrix [64], we add some explicit rules. For example, (i) an attribute name (*e.g.,* **`name`**) can only attend on the other attribute names (*e.g.,* **`expertise`** and **`city`**) and its associated tokens for attribute values (*e.g.,* "Michael" and "Jordan"), but not other attribute values (*e.g.,* "Berkeley"), and (ii) a token for an attribute value (*e.g.,* "Berkeley") can only attend to all attribute values of all attributes and its attribute name (*i.e.,* **`city`**), but not other attribute names (*e.g.,* **`name`**). TURL [19] also uses this technique, called *visibility matrix*.

2.3 Related Work

Data Cleaning. We categorize prior work on data cleaning into three categories. (i) *Only examine the data at hand.* There are integrity constraints (FDs [6], its extensions CFDs [26] and PFDs [54], denial constraints [16], and rule-based methods [33, 66]), and probabilistic based methods (*e.g.,* HoloClean [57]). They need enough signals or data redundancy from D. Supervised ML based methods (*e.g.,* GDR [70], SCAREd [69], Raha [49] and Baran [48]) learn only from the data at hand, which cannot be generalized to other datasets. (ii) *Use External Reliable Sources:* This includes the of master data [28, 37] or knowledge bases [17, 32]. These methods require experts to define or confirm the matching between the data at hand and the external source, *e.g.,* matching rules for table-table matching [28] or graphical patterns for table-graph matching [17]. (iii) *Human- or crowd-in-the-loop.* When neither (i) nor (ii) type solutions work, a last resort is to fall back on humans to clean the dataset.

Intuitively, with enough signals, data redundancy, reliable external sources with sufficient coverage, and the availability of experts to bootstrap and tune the process, we can leverage (i) and (ii) style solutions. Unfortunately, this is usually not the case in practice [1] — cleaning is frequently of type (iii) with its high human cost. Automating type (iii) solutions is the main motivation of RPT.

Knowledge Bases. There are two ways to encode the knowledge: "explicit knowledge" such as knowledge graphs, or "implicit knowledge" by *memorizing* the knowledge using DL models. In practice, both explicit knowledge graphs and implicit pre-trained DL models have been widely studied in industry and academia. Both directions are important, and RPT belongs to the latter. One drawback is that it is hard to explain, for which explainable AI techniques [22, 42, 58] will play an important role for grounding RPT-like tools.

Relational Table Understanding. TaBERT [72], Tapas [35] and TabFact [13] study the question-answering tasks that involve joint reasoning over both free text and structured tables. They take a natural language utterance as input and produce a structured query (*e.g.,* an SQL query in TaBERT or aggregations in Tapas [35]), or a classification result (*e.g.,* support or refute in TabFact). To this end, they focus on learning a joint representation over textual utterances and tabular data with Transformer models and designing various pre-training tasks to this end.

Closer to this work is TURL [19]. However, RPT differs from TURL in two aspects: (i) TURL employs an encoder-only architecture for learned representations, instead of generating a complicated output, *e.g.,* a tuple or a table. The additional decoder architecture of RPT provides the flexibility of generating sequences in multiple forms. (ii) TURL has to be used with a KB to extract values. For example, for cell infilling, TURL uses a pre-trained model (1.2GB) to generate a representation, which has to be linked to the KB (*i.e.,* a

Table 1: Compare RPT with BART (yellow: masked values; green: (partially) correct; pink: wrong).

title	manufacturer	price	Truth	RPT-C	BART
instant home design (jewel case)	topics entertainment	[M]	9.99	9	Topics
disney's 1st & 2nd grade bundle ...	disney	[M]	14.99	19	Dis
adobe after effects professional 6.5 ...	[M]	499.99	adobe	adobe	$1.99
stomp inc recover lost data 2005	[M]	39.95	stomp inc	stomp	39.95
[M]	write brothers	269.99	write brothers dramatica ...	write brothers	1.99

collection of web tables, 4.6GB) to get the actual value. RPT (1.6GB in our experiment) does not need such a KB to fill missing values.

2.4 Opportunities

RPT naturally supports several common data preparation tasks, *e.g.,* error detection, data repairing, auto-completion, auto-suggestion, and schema matching. Yet there are also many opportunities.

(O1) *Hybrid Solutions.* While RPT does not differentiate between categorical or numeric data during pre-training, it works better for categorical data (*i.e.,* human-easy). A promising direction is to combine RPT with other (quantitative) data cleaning methods [53] from a rich set of (a-b) type data cleaning solutions.

(O2) *Dirty Data.* Many tables are dirty. Pre-training RPT on these dirty tables may yield a biased result. Currently, we learn directly from dirty tables, by assuming that the frequency of correct values is higher than the frequency of wrong values. There are several open problems. First, we would like to provide some guarantee of model robustness while still learning from dirty data. Second, a cleaned version of training data that can be used as a benchmark is highly desired, similar to the Colossal Clean Crawled Corpus (C4) for Text-To-Text-Transfer-Transformer (T5) [55].

(O3) *An AI-assisted Tool with Human-in-the-loop.* Achieving high accuracy in diverse data preparation tasks and domains is still a challenge for RPT; it would require substantial in-domain training data. Hence, a practical usage of RPT is to use it as an AI-assisted tool that can suggest meaningful results in many human-in-the-loop tasks, which can guide users and thus reduce human cost.

2.5 Preliminary Result

We have conducted preliminary experiments to show that RPT can reconstruct the masked token(s) in tuples. Our baseline is BART [52], which is pre-trained with a large corpus of text, including from the product domain. Because BART and RPT have the same architecture (Figure 2), we can use the parameters pre-trained by BART, instead of a random initialization. We used tables about products, including Abt-Buy [2] and Walmart-Amazon [65]. Note that these two tables are naturally dirty.

For testing, we used Amazon-Google [3], the tables that were not seen by BART or RPT. We masked attribute values and asked BART and RPT to predict the original values. Table 1 shows some results, where [**M**] means that the value is masked out, column Truth is the ground truth, and columns RPT and BART provide the results predicted by each, respectively. Tuples 1-2 involve predicting missing prices, where RPT gives close predictions but BART does not. Tuples 3-4 involve predicting missing manufacturers; RPT-C provides good predictions. Tuple 5 involves predicting a missing title, and RPT provides a partially correct prediction.

This preliminary experiment shows that RPT pre-trained on tables can learn structural data values from tables better than directly using a pre-trained language model (*e.g.,* BART), which is not customized for relational data. The main reason is that, by pre-training on the product tables, RPT can better learn dependency among columns, and thus is more capable of predicting missing values. Of course, RPT sometimes makes wrong predictions, but for those cases, BART also fails. Our belief is that these preliminary results are suggestive enough of the effectiveness of the approach that it merits significant additional investigation.

Limitations. RPT faces similar limitations that pre-trained language models (LMs) face. (1) *Numeric values:* numeric values are usually mapped into unknown tokens causing the model to fail on tasks that require precise prediction on numeric values. (2) *Max sequence length:* restricted by the GPU memory size, most pre-trained LMs are limited by the sequence length, thus data preparation on wide tables may require additional optimization. (3) *Not fully reliable.* Similar to GPT-3, a generative model cannot be fully trusted. One way to combat this is to treat it as an AI-assistant with a human-in-the-loop, as discussed in Section 2.4 Opportunities (O3).

3 FINE-TUNING RPT

The encoder-decoder architecture of RPT (pre-trained on *tuple-to-tuple*) provides the flexibility to be fine-tuned for different downstream data preparation tasks (*i.e., tuple-to-X*).

Value Normalization. Because RPT has an autoregressive decoder, it can be directly fine-tuned for sequence generation tasks such as value normalization (*e.g.,* "Mike Jordan, 9 ST, Berkeley" → "Mike Jordan, 9th Street, Berkeley"). The encoder takes the input value as a sequence and the decoder generates the output autoregressively. In addition, normalizing "Mike" to "Michael" or "Sam" to "Samuel" can be fine-tuned as a neural name translation [63] task.

Data Transformation. Similarly to what is described above, RPT can be fine-tuned for transformation of data from one format (*e.g.,* a tuple) to another format (*e.g.,* JSON or XML), where the decoder will autoregressively serialize the output in the target format.

Data Annotation. Given a tuple, data annotation requires adding a label (*e.g.,* a classification task). We can use the final hidden state of the final decoder token to fine-tune a multi-class linear classifier.
Information Extraction (IE). Given a tuple, IE extracts a span or multiple spans of relevant text, which can be done by fine-tuning the decoder to produce the (start, end) pairs of spans.

Learned Tuple Representation for Entity Resolution. The embeddings of entities have been used in entity resolution for both blocking [23] and entity matching [23, 50]. A typical trick is to

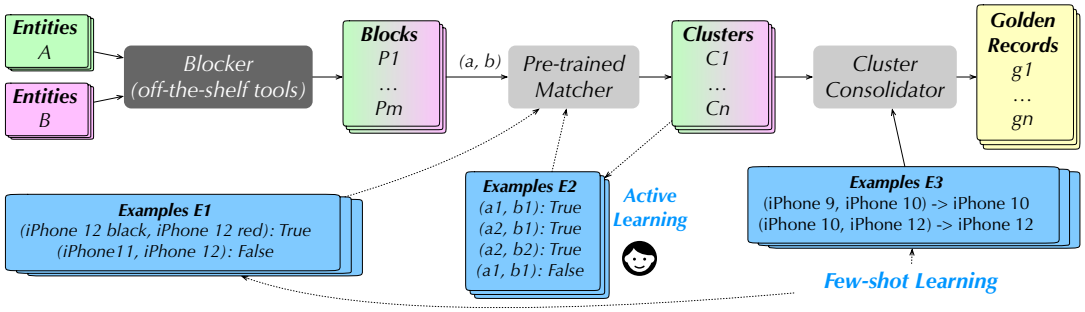

Figure 3: Collaborative Learning and Few-shot Learning for Entity Resolution.

do cross-tuple training (or contrastive learning [12]), via Siamese NNs [14], such that similar entities have similar embeddings. Similarly, the encoder of RPT can be fine-tuned in Siamese NNs for learned representations *w.r.t.* entity resolution.

4 BEYOND RPT

In this section, we explore other, but related, techniques that can help on specific tasks.

4.1 Entity Resolution

Given two sets of entities, A and B, an end-to-end of ***entity resolution pipeline*** (Figure 3) is: (1) find duplicated entity pairs ($a \in A, b \in B$) (blocking to improve efficiency); (2) merge them into clusters, typically through transitive closure, and (3) consolidate each cluster into one entity.

Blocking. There is a rich literature on automatic blocking for ER (see [51] for a survey). There are also DL-based methods [23, 45, 67] to generate blocks. These prior works are automatic and already work well, hence will not be covered in this paper.

Matcher. The state-of-the-art matchers are all ML based *e.g.*, random forests (*e.g.,* Magellan [43]), or DL based (*e.g.,* DeepMatcher [50] and DeepER [23]). Recent works [9, 45] also study to leverage pre-trained LM models for generating entity representations.

Consolidator. There are rule-based methods [27] and learning-based approach [34] for entity consolidation – both need either significant human involvement or a large amount of training data.

Vision and Opportunities. The Matcher and Consolidator should be able to perform effectively through pre-trained models (*i.e.,* to obtain knowledge) and a few examples (*i.e.,* to interpret the task).

However, this pipeline cannot be fully automated, because some judgments are *objective*, *e.g.,* "iPhone 10", "iPhone ten", and "iPhone X" are the same, while some others are *subjective*, *e.g.,* whether "iPhone 12 red" matches "iPhone 12 black" is user dependent.

Our intuition is that the objective criteria can be pre-trained (such as "iPhone 10" matches "iPhone X" and "Mike" matches "Michael"), but the subjective criteria need task-specific samples, for both the Matcher and the Consolidator (Figure 3), which could be achieved by getting a few examples from humans.

We identify two major opportunities for the entire ER pipeline.

(O1) *Collaborative learning* or *Federated Learning (FL) [7, 71].* This is to learn the "objective" criteria for the Matcher. Note that there are many public and private ER benchmarks, which share common domains. It is promising to collaboratively train one Matcher, and the knowledge can be learned and transferred from one dataset to another dataset. Better still, this can be done securely [47], without data sharing. Note that, there have been transfer learning techniques on ER [41, 61] to show an early success on this thread.

We believe that we should build a platform collaboratively for ER, with a pre-trained model M for each domain. Anyone who wants to benefit from M can download M, retrain using his/her data to get a M_1, and send back an update of parameters $\Delta_1 = M_1 - M$, and the platform will merge the model update with M, from multiple users [7]. Because different entities may have different schemas, we use a pre-trained model such as BERT to be schema-agnostic.

(O2) *Few-shot Learning.* This is to learn the "subjective" criteria, for Matcher and Consolidator, through a human-in-the-loop approach. The goal is to infer a better specified task from a few examples, *e.g.,* using Pattern-Exploiting Training [59].

[Matcher.] Consider E_1 in Fig. 3 that contains two user provided examples and we want to automatically generate a clearer task for workers, *e.g.,* "color does not matter but model matters". We can design two templates like (T1) "True: if $\boxed{a}$ and $\boxed{b}$ have the same $[\mathbf{M}]_1$" and (T2) "False: if $\boxed{a}$ and $\boxed{b}$ have different $[\mathbf{M}]_2$". By replacing the first matching pair in E_1 to template (T1), we can infer a pattern "model" or "series" (but not "color") for $[\mathbf{M}]_1$. Similarly, by using the second un-matching pair in E_2 to template (T2), we can infer a pattern "model" or "series" for $[\mathbf{M}]_2$.

Moreover, when merging matching entities into clusters based on transitive closure, conflict may be automatically detected within clusters (*e.g.,* E_2 in Fig. 3); such conflicts can be resolved by the users through active learning. Note that, doing active learning from conflicting predictions is different from traditional active learning methods on ER that use confusing/informative entity pairs [4, 15].

[Consolidator.] Consider E_3 with two examples, "iPhone 10 is more preferred than iPhone 9", and "iPhone 12 is more preferred than iPhone 10". We can use them to make the task clearer by asking questions "iPhone 10 is [**M**] than iPhone 9" and "iPhone 12 is [**M**] than iPhone 10", and enforce a language model to fill the two masked tokens with the same value, which might be "newer".

Another powerful method related to few-shot learning is meta-learning (or learning to learn fast) [62], with the main goal to learn new concepts and skills fast with a few examples.

Preliminary Results. We have conducted some preliminary experiments on the "product" domain for the Matcher, because this domain has a rich collection of ER benchmarks, and it is

Table 2: Comparison with the State of the art.

Method	Abt-Buy		Amazon-Google	
	F1 score	# labels	F1 score	# labels
Collaborative Training (CT)	0.72	0	0.53	0
ZeroER	0.52	0	0.48	0
DeepMatcher	0.63	7689	0.69	9167
Ditto	0.71	1500	0.50	1000

known to be hard ER cases because they are text-rich. Specifically, we use five well-known benchmarks. (D1) Abt-Buy [2], (D2) Amazon-Google [3], (D3) Walmart-Amazon [65], (D4) iTunes-Amazon [38], and (D5) SIGMOD 2020 programming contest [60] (we took 1000/8000 matching/unmatching pairs).

Specifically, when testing on D1, we train with D2–D5, and when testing on D2, we train with D1, D3–D5. We only tested on D1 and D2, so we can directly compare with ZeroER [68] (without any training data) and DeepMatcher [50] (trained with hundreds/thousands of examples), where the F1 scores and the number of labels are reported from their original papers. We also compare with Ditto [45]. We progressively add the number of labeled data to fine-tune Ditto until its F1 is very close to that of collaborative training (CT).

We also note that the paper [9] is quite similar to Ditto. First, both studies use the same strategy to model a record pair as a sequence and fine-tune a pre-trained model to output the matching results. Second, [9] compares different pre-trained LM models for entity resolution and reports that RoBERTa [23] achieves the best performance. Similarly, the latest version of Ditto [45] also uses RoBERTa as its pre-trained model. Third, both studies achieve similar results in the experiments. For example, on the Abt-Buy dataset, [9] and Ditto achieve 0.91 and 0.90 on F1 score respectively.

Table 2 shows that CT outperforms ZeroER and is comparable with DeepMatcher that was trained with 1000+ examples. Moreover, CT uses zero examples from the test ER dataset to achieve the performance of Ditto, which is trained by fine-tuning a pre-trained model with 1000+ examples. This result verifies the opportunity (O1) that it is promising to collaboratively train a Matcher to decide whether two entities (even in different schemas) match or not.

4.2 Information Extraction

Information Extraction (IE) is the process of retrieving specific information from unstructured (*e.g.,* text), or (semi-)structured (*e.g.,* relational) data. We consider a simple IE problem: given a text or text-rich tuple t, it is to extract a span (*i.e.,* a continuous sequence of tokens) from t, denoted by $\mathbf{I}(t)$. See Figure 1(c) for a sample IE task. Although simple, the above definition is general enough to cover a wide range of IE problems.

Connection with Question-Answering in NLP. There is a natural connection between the IE problem and a typical question-answering problem in NLP. For question answering, there is an input question such as "Q: where do water droplets collide with ice crystals to form precipitation", and an input paragraph "P: ... Precipitation forms as smaller droplets coalesce via collision with other rain drop or ice crystals within a cloud. ...". The task is to find a span *e.g.,* within a cloud of paragraph P to answer the question Q. There are many NLP question-answering benchmarks, such as SQuAD [56] where AI has outperformed human performance [73].

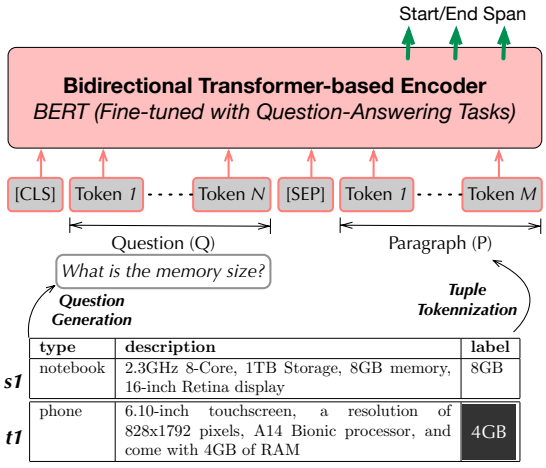

Figure 4: Connecting IE with NLP Question-Answering.

As shown in Figure 4, given a query Q and a paragraph P, a pre-trained model fine-tuned using question-answering benchmarks can provide a span, *i.e.,* (start, end) positions, as output.

We can tokenize a tuple t as a paragraph P (Section 2). The remaining problem is to generate the question Q. We can have a question template such as "what is the [**M**]", where the [**M**] can be instantiated with one-shot learning (*e.g.,* the label of s_1) via *e.g.,* PET [59], which gives "what is the memory size" as the question Q.

Opportunities. (O1) Connect more DB-related IE tasks to well-studied NLP tasks, so as to obtain pre-trained knowledge (*e.g.,* NeruON [5] uses a seq-to-seq model to extract tuples from question-answer pairs). (O2) Currently, many IE tasks are performed by crowd workers (or crowd-in-the-loop). Instead of fully replacing these crowd workers, we are studying how to train multiple RPT-I models as AI-workers, and mix the AI-workers and crowd workers to reduce the total cost of a crowdsourcing task.

5 CALL TO ARMS

We have presented our vision and concrete steps for democratizing data preparation: RPT, fine-tuning RPT, and other appealing techniques. Several recent successes (*e.g.,* Termite [29], EmbDI [11], TURL [19], Ditto [45] and NeurON [5]) have shed some light on this direction. Our preliminary results, along with these related papers suggest that learning-based approaches have the potential to outperform more traditional methods, much as they have revolutionized NLP. However, the data preparation field is vast, the problems are diverse and much work remains to be done. In particular, a major obstacle to advance all the above topics is the limited availability of real-world benchmarks, *e.g.,* C4 for T5 [55]. Now is the time for the data preparation and larger database communities to come together to explore the potential of these new techniques.

ACKNOWLEDGMENTS

This work was partly supported by National Key Research and Development Program of China (2020YFB2104101), NSF of China (61632016, 61925205, 62072461, U1911203), Huawei, TAL Education, and Beijing National Research Center for Information Science and Technology.

REFERENCES

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting Data Errors: Where are we and what needs to be done? *Proc. VLDB Endow.* 9, 12 (2016), 993–1004.

[2] Abt-Buy. [n.d.]. https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#abt-buy.

[3] Amazon-Google. [n.d.]. https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#amazon-google.

[4] Kedar Bellare, Suresh Iyengar, Aditya G. Parameswaran, and Vibhor Rastogi. 2013. Active Sampling for Entity Matching with Guarantees. *ACM Trans. Knowl. Discov. Data* 7, 3 (2013), 12:1–12:24.

[5] Nikita Bhutani, Yoshihiko Suhara, Wang-Chiew Tan, Alon Y. Halevy, and H. V. Jagadish. 2019. Open Information Extraction from Question-Answer Pairs. In *NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). 2294–2305.

[6] Philip Bohannon, Michael Flaster, Wenfei Fan, and Rajeev Rastogi. 2005. A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification. In *SIGMOD*, Fatma Özcan (Ed.). 143–154.

[7] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. 2019. Towards Federated Learning at Scale: System Design. In *MLSys*.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[9] Ursin Brunner and Kurt Stockinger. 2020. Entity Matching with Transformer Architectures - A Step Forward in Data Integration. In *EDBT*. 463–473.

[10] Michael J. Cafarella, Jayant Madhavan, and Alon Y. Halevy. 2008. Web-scale extraction of structured data. *SIGMOD Rec.* 37, 4 (2008), 55–61.

[11] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. 2020. Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. In *SIGMOD*. 1335–1349.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.

[13] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. TabFact: A Large-scale Dataset for Table-based Fact Verification. *CoRR* abs/1909.02164 (2019).

[14] Davide Chicco. 2021. Siamese Neural Networks: An Overview. In *Artificial Neural Networks - Third Edition*. Methods in Molecular Biology, Vol. 2190. 73–94.

[15] Victor Christen, Peter Christen, and Erhard Rahm. 2019. Informativeness-Based Active Learning for Entity Resolution. In *ECML PKDD*, Peggy Cellier and Kurt Driessens (Eds.), Vol. 1168. 125–141.

[16] Xu Chu, Ihab F. Ilyas, and Paolo Papotti. 2013. Discovering Denial Constraints. *Proc. VLDB Endow.* 6, 13 (2013), 1498–1509.

[17] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. 2015. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In *SIGMOD*. 1247–1261.

[18] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. 2017. The Data Civilizer System. In *CIDR*.

[19] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. TURL: Table Understanding through Representation Learning. *Proc. VLDB Endow.* (2020).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. [n.d.]. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). 4171–4186.

[21] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. 2012. *Principles of Data Integration*. Morgan Kaufmann.

[22] Finale Doshi-Velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017).

[23] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed Representations of Tuples for Entity Resolution. *Proc. VLDB Endow.* 11, 11 (2018), 1454–1467.

[24] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* 19, 1 (2007), 1–16.

[25] Wenfei Fan and Floris Geerts. 2012. *Foundations of Data Quality Management*. Morgan & Claypool Publishers.

[26] Wenfei Fan, Floris Geerts, Xibei Jia, and Anastasios Kementsietsidis. 2008. Conditional functional dependencies for capturing data inconsistencies. *ACM Trans. Database Syst.* 33, 2 (2008), 6:1–6:48.

[27] Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. 2013. Inferring data currency and consistency for conflict resolution. In *ICDE*, Christian S. Jensen, Christopher M. Jermaine, and Xiaofang Zhou (Eds.). 470–481.

[28] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Wenyuan Yu. 2010. Towards Certain Fixes with Editing Rules and Master Data. *Proc. VLDB Endow.* 3, 1 (2010), 173–184.

[29] Raul Castro Fernandez and Samuel Madden. 2019. Termite: a system for tunneling through heterogeneous data. In *Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, aiDM@SIGMOD 2019, Amsterdam, The Netherlands, July 5, 2019*. 7:1–7:8.

[30] Behzad Golshan, Alon Y. Halevy, George A. Mihaila, and Wang-Chiew Tan. 2017. Data Integration: After the Teenage Years. In *PODS*, Emanuel Sallinger, Jan Van den Bussche, and Floris Geerts (Eds.). ACM, 101–106.

[31] Mazhar Hameed and Felix Naumann. 2020. Data Preparation: A Survey of Commercial Tools. *SIGMOD Rec.* 49, 3 (2020), 18–29.

[32] Shuang Hao, Nan Tang, Guoliang Li, and Jian Li. 2017. Cleaning Relations Using Knowledge Bases. In *ICDE*. 933–944.

[33] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. 2016. Interactive and Deterministic Data Cleaning. In *SIGMOD*. 893–907.

[34] Alireza Heidari, George Michalopoulos, Shrinu Kushagra, Ihab F. Ilyas, and Theodoros Rekatsinas. 2020. Record fusion: A learning approach. *CoRR* abs/2006.10208 (2020).

[35] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). 4320–4333.

[36] Ihab F. Ilyas and Xu Chu. 2019. *Data Cleaning*. ACM.

[37] Matteo Interlandi and Nan Tang. 2015. Proof positive and negative in data cleaning. In *ICDE*. 18–29.

[38] Itunes-Amazon. [n.d.]. https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#itunes-amazon.

[39] Ayush Jain, Akash Das Sarma, Aditya G. Parameswaran, and Jennifer Widom. 2017. Understanding Workers, Developing Effective Tasks, and Enhancing Marketplace Dynamics: A Study of a Large Crowdsourcing Marketplace. *Proc. VLDB Endow.* 10, 7 (2017), 829–840.

[40] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics* 8 (2020), 64–77.

[41] Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource Deep Entity Resolution with Transfer and Active Learning. In *ACL*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). 5851–5861.

[42] Been Kim and Finale Doshi-Velez. 2017. Interpretable Machine Learning: The fuss, the concrete and the questions. In *ICML Tutorial*.

[43] Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, Adel Ardalan, Jeffrey R. Ballard, Han Li, Fatemah Panahi, Haojun Zhang, Jeffrey F. Naughton, Shishir Prasad, Ganesh Krishnan, Rohit Deep, and Vijay Raghavendra. 2016. Magellan: Toward Building Entity Matching Management Systems. *Proc. VLDB Endow.* 9, 12 (2016), 1197–1208.

[44] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*. 7871–7880.

[45] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. [n.d.]. Deep Entity Matching with Pre-Trained Language Models. *Proc. VLDB Endow.* ([n. d.]).

[46] Bin Liu, Laura Chiticariu, Vivian Chu, H. V. Jagadish, and Frederick Reiss. 2010. Automatic Rule Refinement for Information Extraction. *Proc. VLDB Endow.* 3, 1 (2010), 588–597.

[47] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. 2020. A Secure Federated Transfer Learning Framework. *IEEE Intell. Syst.* 35, 4 (2020), 70–82.

[48] Mohammad Mahdavi and Ziawasch Abedjan. 2020. Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning. *Proc. VLDB Endow.* 13, 11 (2020), 1948–1961.

[49] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. 2019. Raha: A Configuration-Free Error Detection System. In *SIGMOD*. 865–882.

[50] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *SIGMOD*, Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein (Eds.). 19–34.

[51] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. 2020. Blocking and Filtering Techniques for Entity Resolution: A Survey. *ACM Comput. Surv.* 53, 2 (2020), 31:1–31:42.

[52] Ofir Press, Noah A. Smith, and Omer Levy. [n.d.]. Improving Transformer Models by Reordering their Sublayers. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). 2996–3005.

[53] Nataliya Prokoshyna, Jaroslaw Szlichta, Fei Chiang, Renée J. Miller, and Divesh Srivastava. 2015. Combining Quantitative and Logical Data Cleaning. *Proc. VLDB Endow.* 9, 4 (2015), 300–311.

[54] Abdulhakim Ali Qahtan, Nan Tang, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. 2020. Pattern Functional Dependencies for Data Cleaning. *Proc. VLDB Endow.* 13, 5 (2020), 684–697.

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67.

[56] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). 2383–2392.

[57] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proc. VLDB Endow.* 10, 11 (2017), 1190–1201.

[58] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *SIGKDD*. 1135–1144.

[59] Timo Schick and Hinrich Schütze. 2020. Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference. *CoRR* abs/2001.07676 (2020).

[60] sigmod-2020-contest. [n.d.]. http://www.inf.uniroma3.it/db/sigmod2020contest/task.html.

[61] Saravanan Thirumuruganathan, Shameem Ahamed Puthiya Parambath, Mourad Ouzzani, Nan Tang, and Shafiq R. Joty. 2018. Reuse and Adaptation for Entity Resolution through Transfer Learning. *CoRR* abs/1809.11084 (2018).

[62] Sebastian Thrun and Lorien Y. Pratt. 1998. Learning to Learn: Introduction and Overview. In *Learning to Learn*, Sebastian Thrun and Lorien Y. Pratt (Eds.). Springer, 3–17.

[63] Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural Machine Translation Incorporating Named Entity. In *COLING*. 3240–3250.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[65] Walmart-Amazon. [n.d.]. https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#walmart-amazon.

[66] Jiannan Wang and Nan Tang. 2014. Towards dependable data repairing with fixing rules. In *SIGMOD*. ACM, 457–468.

[67] Zhengyang Wang, Bunyamin Sisman, Hao Wei, Xin Luna Dong, and Shuiwang Ji. 2020. CorDEL: A Contrastive Deep Learning Approach for Entity Linkage. *ICDM*.

[68] Renzhi Wu, Sanya Chaba, Saurabh Sawlani, Xu Chu, and Saravanan Thirumuruganathan. 2020. ZeroER: Entity Resolution using Zero Labeled Examples. In *SIGMOD*. 1149–1164.

[69] Mohamed Yakout, Laure Berti-Équille, and Ahmed K. Elmagarmid. 2013. Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes. In *SIGMOD*, Kenneth A. Ross, Divesh Srivastava, and Dimitris Papadias (Eds.). ACM, 553–564.

[70] Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F. Ilyas. 2011. Guided data repair. *Proc. VLDB Endow.* 4, 5 (2011), 279–289.

[71] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2 (2019), 12:1–12:19.

[72] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *ACL*. 8413–8426.

[73] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective Reader for Machine Reading Comprehension. arXiv:cs.CL/2001.09694