Attention-enhanced Gradual Machine Learning for Entity Resolution

Ping Zhong
School of Computer Science, Northwestern Polytechnical University

Zhanhuai Li
School of Computer Science, Northwestern Polytechnical University

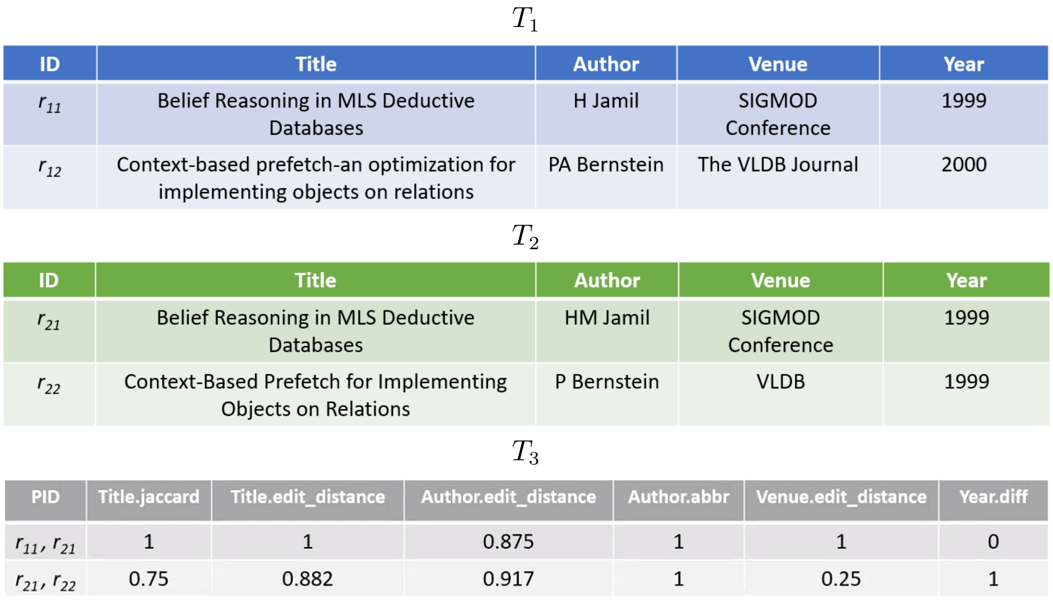
Qun Chen
School of Computer Science, Northwestern Polytechnical University

Boyi Hou
School of Computer Science, Northwestern Polytechnical University

Abstract—
Recent work has shown that Entity Resolution (ER) can be effectively performed by Gradual Machine Learning (GML). GML begins with some automatically labeled easy instances, and then gradually labels more challenging instances by iterative factor graph inference without human intervention. In GML, shared features serve as the medium for knowledge conveyance between easy instances and more challenging ones. The existing GML solution supposes that features play independent roles in gradual inference. However, in real scenarios, this assumption may be untenable since features are usually correlated with each other. To address this limitation, this paper proposes an attention-enhanced approach to improve the accuracy of gradual inference. We first propose a method of spectral feature representation to map correlated features to close points in the same vector space, and then present a model of attention neural network to learn the decisive features given arbitrary combinations of features for improved feature weighting. Finally, our extensive experiments on real benchmark data have validated the efficacy of the proposed approach.

■ INTRODUCTION

The task of Entity resolution (ER) aims to identify equivalent records that refer to the same real-world entity. Consider the running example shown in Figure 1. ER needs to match the paper records between two tables, T_1 and T_2. The pair of $< r_{1i}, r_{2j} >$, in which r_{1i} and r_{2j} denote a record in T_1 and T_2 respectively, is called an *equivalent* pair if and only if r_{1i} and r_{2j} refer to the same paper; otherwise, it is called an



T_1

ID	Title	Author	Venue	Year
r_{11}	Belief Reasoning in MLS Deductive Databases	H Jamil	SIGMOD Conference	1999
r_{12}	Context-based prefetch-an optimization for implementing objects on relations	PA Bernstein	The VLDB Journal	2000

T_2

ID	Title	Author	Venue	Year
r_{21}	Belief Reasoning in MLS Deductive Databases	HM Jamil	SIGMOD Conference	1999
r_{22}	Context-Based Prefetch for Implementing Objects on Relations	P Bernstein	VLDB	1999

T_3

PID	Title.jaccard	Title.edit_distance	Author.edit_distance	Author.abbr	Venue.edit_distance	Year.diff
r_{11}, r_{21}	1	1	0.875	1	1	0
r_{21}, r_{22}	0.75	0.882	0.917	1	0.25	1

Figure 1: An ER Example: (1) T_1: record table 1; (2) T_2: record table 2; (3) T_3: the table of feature set.

inequivalent pair. In the example, r_{11} and r_{21} are *equivalent* while r_{11} and r_{22} are *inequivalent*.

The state-of-the-art performance on ER has been achieved by deep learning [1], [2]. However, the efficacy of these DNN models depends on the presence of a large quantity of accurately labeled training data, which may not be readily available in real scenarios. To alleviate this limitation, recent work has shown that ER can be effectively performed by Gradual Machine Learning (GML) [3], [4], which can enable automatic machine labeling without manual intervention. Given a classification task, GML begins with some easy instances, which can usually be automatically labeled by the machine with high accuracy, and then gradually labels the more challenging instances by iterative factor graph inference. The following two properties of GML make it fundamentally different from the existing learning paradigms:

- Distribution misalignment between easy and hard instances in a task. GML processes the instances in the increasing order of hardness. Its scenario does not satisfy the i.i.d (independent and identically distributed) assumption underlying most existing machine learning models: the labeled easy instances are not representative of the unlabeled harder ones.
- Gradual learning by small stages. At each stage, GML typically labels only one instance based on the evidential certainty provided by labeled easier instances. The process of iterative labeling can be performed in an unsupervised manner without requiring any human intervention.

However, the existing GML solution for ER supposes that features play independent roles in gradual inference. Unfortunately, this assumption may be untenable in real scenarios because features are usually correlated with each other. Consider the running example shown in Fig 1, in which T_3 lists the metric features leveraged for gradual inference. It can be observed that the records, r_{11} and r_{21}, are highly similar on all the metrics. Therefore, the pair $< r_{11}, r_{21} >$ could be reasoned to be equivalent. However, it is interesting to note that r_{12} and r_{22} are also similar on most metrics except that they have different publication years. If the metrics are treated as independent features, it is very likely that the pair $< r_{12}, r_{22} >$ would also be reasoned to be equivalent. However, in this case, the influence of other metrics is to a large extent dependent on the metric value of *Year.diff*. It can be observed that r_{12} is indeed a follow-up work of r_{22} and they represent different publications.

Therefore, there is a need for feature correlation analysis to improve feature influence estimation. In this paper, we propose a novel approach of attention-enhanced gradual inference for GML in this paper. Widely used in various natural language processing tasks (e.g. auto translation [5]), the existing attention neural networks usually leverage pre-trained models such as BERT to represent text as fixed-length vectors. These models map semantically similar text to close points in the same vector space. However, the scenario of GML brings about new challenges because GML requires to map features with similar distributions to close points in the same vector space. Our main contributions can be summarized as follows:

- We propose a novel approach of attention-enhanced gradual inference for GML, which can automatically optimize feature weighting by feature correlation analysis;
- We present a model of attention neural network to enable attention-enhanced gradual inference. In particular, we present a new method of spectral feature representation to map features with similar distributions into the same vector space. Based on spectral feature representation, the proposed model can effectively learn the decisive features given arbitrary combinations of features.
- We empirically validate the efficacy of the proposed approach on real benchmark data. Our extensive experiments have shown that the proposed approach considerably outperforms the existing GML as well as other unsupervised alternatives, and its performance is even competitive with the supervised DNN solutions.

Related Work

ER plays a key role in data integration and has been extensively studied in the literature [6]. The state-of-the-art performance on ER has been achieved by deep learning [1], [2]. However, the efficacy of these DNN models depends on the presence of a large quantity of accurately labeled training data, which may not be readily available in real scenarios.

GML has also been applied to the task of aspect-based sentiment analysis [7]. Curriculum learning (CL) [8] and self-paced learning (SPL) [9] are to some extent similar to gradual machine learning in that they generally start with learning easier aspects of a task, and then gradually takes more complex examples into consideration. However, the models trained by curriculum learning or self-paced learning are supposed to be applied on a target workload satisfying the i.i.d assumption. Therefore, as traditional supervised learning, their efficacy still depends on good-quality training examples.

Attention neural networks have been widely used to discriminate relevant information in image inpainting [10] and natural language processing tasks [5]. In particular, a special multistage attention module was proposed to restore the mask regions of damaged images in [10]. It is noteworthy that the existing attention mechanisms can leverage readily available feature representations (e.g. image and pre-trained language model representations). In contrast, this paper proposes a novel approach based on spectral embedding to represent features with different distributions as vectors such that they can be processed by an attention neural network for correlation analysis.

GML Paradigm Overview

The process of gradual machine learning for ER, as shown in Figure 2 , consists of the following three essential steps:

Easy Instance Labeling

For ER, it can be observed that the more similar two records are, the more likely they refer to the same real-world entity. According to this assumption, we can *statistically* state that a pair with a high (resp. low) similarity has a correspondingly high probability of being an equivalent (resp. inequivalent) pair. These record pairs can be deemed to be easy in that they can be automatically labeled by the machine with high accuracy.

Feature Modeling

To facilitate effective knowledge conveyance between labeled and unlabeled instances, we extract the following two types of features from ER record pairs:

1) Attribute value similarity. This type of feature measures a pair's value similarity at

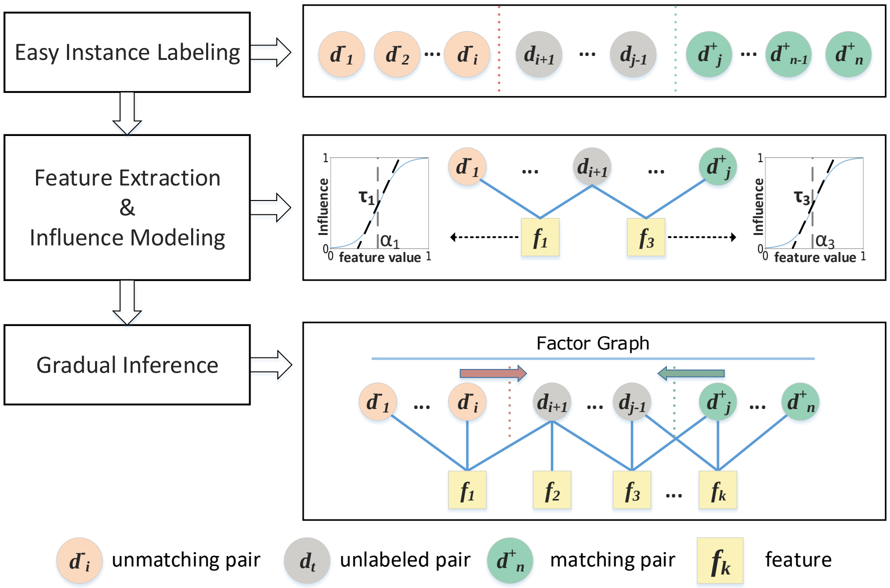


Figure 2: Paradigm Overview. The GML consists of the following three essential steps:(1) Easy Instance Labeling; (2)Feature Extracting and Influence Modeling; (3) Gradual Inference.

each record attribute. It is noteworthy that different attributes usually require different similarity metrics.

2) Token features. Denoting a token by o_i, we represent the feature that o_i occurs in both records by $Same(o_i)$, and the feature that o_i occurs in one and only one record by $Diff(o_i)$. Note that the feature of $Same(o_i)$ serves as evidence for *equivalence*, while $Diff(o_i)$ indicates *inequivalence*. For the workloads with miscellaneous tokens, not every token is highly discriminating (or indicative of entity identity); therefore, we filter the tokens by the metric of IDF (inverse document frequency).

For each extracted feature, GML models its influence over the labels of its relevant instances by a monotonous sigmoid function with two parameters, α and τ, which denote the x-value of the function's midpoint and the steepness of the curve respectively, as follows

$$P_f(d) = \frac{1}{1 + e^{-\tau_f (x_f(d) - \alpha_f)}}, \tag{1}$$

in which f denotes a feature, d denotes pair instance, $x_f(d)$ denotes d's feature value w.r.t f and $P_f(d)$ denotes the influence of f over d. Note that token features have the constant value of 1. Therefore, we first align them with record similarity and then model their influence by sigmoid functions.

Gradual Inference.

GML construct a factor graph, G, which consists of the variables representing labeled instances and unlabeled harder ones, and the factors representing the common features between instances. Denoting the feature set of a pair d by F_d, a factor graph infers the equivalence probability of d, $P(d)$, by:

$$P(d) = \frac{\prod_{f \in F_d} e^{\omega_f(d)}}{1 + \prod_{f \in F_d} e^{\omega_f(d)}}, \tag{2}$$

where

$$\begin{aligned} \omega_f(d) &= \theta_f(d) \cdot \log\left(\frac{P_f(d)}{1 - P_f(d)}\right) \\ &= \theta_f(d) \cdot \tau_f(x_f(d) - \alpha_f), \end{aligned} \tag{3}$$

in which $\omega_f(d)$ denotes the factor weight of f, $\log(\cdot)$ encodes the estimated influence of f on d by sigmoid regression, and $\theta_f(d)$ represents the confidence on influence estimation.

In GML, the parameters of (α,τ) need to be iteratively learned by maximum likelihood. A scalable approach of gradual inference has also been presented in [4]. It first selects the top-m unlabeled variables with the most evidential support in G as the candidates for probability inference. To reduce the invocation of maximum likelihood estimation, it then approximates probability inference by an efficient algorithm on the m candidates. Finally, it estimates the probabilities of only the top-k most promising unlabeled variables among the m candidates via factor graph inference. At each iteration, GML labels the unlabeled pair with the highest degree of evidential certainty measured by inverse of entropy.

Attention-enhanced Gradual Inference

In this section, we first present the method of spectral feature representation, and then describe the attention model to enable attention-enhanced gradual inference.

Spectral Feature Representation

To capture feature correlation, we first construct a distribution similarity matrix of features, and then generate their corresponding spectral embeddings, which are finally taken as desired feature representations. We measure the similarity of two feature distributions based on their co-occurrence in pair instances. Specifically, we use the metric of Maximal Information Coefficient (MIC) because it has been empirically shown to be effective at capturing correlations among distributions with diversified shapes.

Formally, we denote a feature set by F and its matrix of $|F|\times|F|$ by W, in which each entrance w_{ij} is estimated by

$$w_{ij} = \frac{|D_{f_i} \cap D_{f_j}|}{|D_{f_i} \cup D_{f_j}|} MIC(f_i, f_j), \tag{4}$$

where f_i denotes a feature in F, D_{f_i} the set of instances having the feature f_i, and $MIC(f_i, f_j)$ the Maximal Information Coefficient between f_i and f_j. Specifically, the value of $MIC(f_i, f_j)$ is estimated by

$$MIC(f_i, f_j) = \max_{r_i \times r_j < B} \frac{\max\limits_{G \in \mathbb{G}_{r_i \times r_j}} (I_G)}{log\, min(r_i, r_j)}, \tag{5}$$

in which B denotes a function of sample size, $\mathbb{G}_{r_i \times r_j}$ denotes all grids on the scatter plot of the two variables having the resolution size of $r_i \times r_j$, and I_G denotes the mutual information of the probability distribution induced in the box of G.

To extract spectral embeddings based on the matrix of W, we compute the normalized Graph Laplacians L for the matrix W, which is formally represented by

$$L = I - \tilde{D}^{\frac{1}{2}} W \tilde{D}^{\frac{1}{2}}, \tag{6}$$

where I denotes the identity matrix and $\tilde{D}$ the degree matrix. Specifically, $\tilde{D}$ is a diagonal matrix with $\tilde{d}_1, \cdots, \tilde{d}_{|F|}$ on its diagonal, and the value of $\tilde{d}_i$ is computed by

$$\tilde{d}_i = \sum_{j=1}^{|F|} w_{ij}. \tag{7}$$

Subsequently, we compute the first m eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_m$ of L, where each $\mathbf{v}_i$ denotes a vector with the size of $|F|$. Let $\mathbf{e} \in \mathbb{R}^{|F|\times m}$ denote a matrix consisting of the vectors $\mathbf{v}_1, \cdots, \mathbf{v}_m$. Accordingly, the spectral embedding $\mathbf{e}_i \in E$ of a feature, f_i, is represented by

$$\mathbf{e}_i = (\mathbf{v}_1^i, \cdots, \mathbf{v}_m^i). \tag{8}$$

Finally, the attention-enhanced GML represents each feature, f_i, by its spectral embedding of $\mathbf{e}_i$, which is taken as the input to attention neural network.

Attention Model

The attention model is supposed to automatically optimize weighting of features by capturing their correlation. For simplicity of presentation, we denote the original estimated factor weight by $\omega_f(d)$, and the new attention-enhanced factor weight by $\phi_f(d)$. As shown in Fig 3, the attention model receives its inputs from two sources: the original factor weights estimated by Eq. 3 and the feature representations as shown in Eq. 8. It outputs a new feature weight for each of the factors in G.

The attention model consists of two parts: an attention multi-layer and a linear transformation layer. It employs the element-wise product of spectral embeddings, $\mathbf{e}_i$, and factor weights, $\omega_{f_i(d)}$, as its inputs. Formally, given an instance d and its feature f_i, its corresponding input, $\mathbf{u}_i^0$,

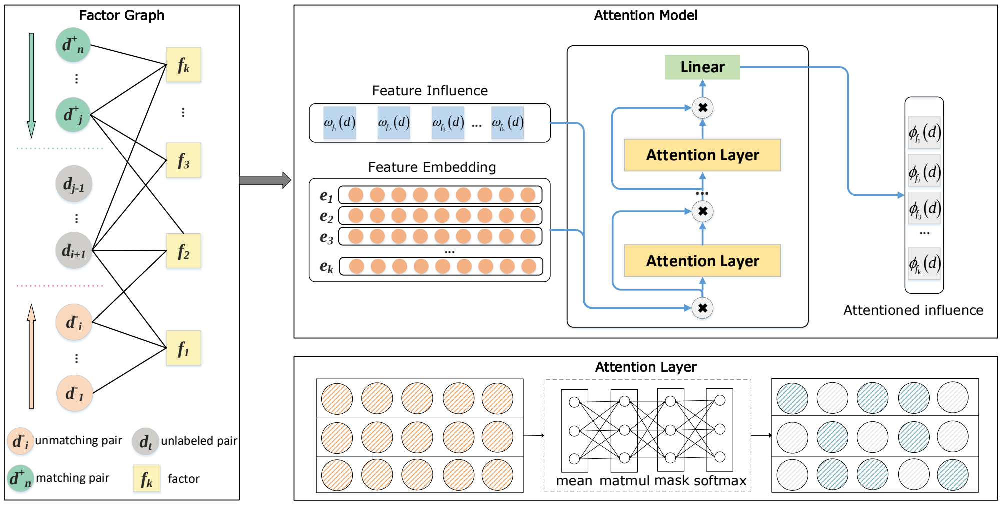


Figure 3: Attentional Gradual Inference. The model converts the estimated factor weight $\omega_f(d)$ to the attention-enhanced factor weight $\phi_f(d)$.

is represented by

$$\mathbf{u}_i^0 = \omega_{f_i}(d) * \mathbf{e_i}, \tag{9}$$

in which $\mathbf{e_i}$ denotes the embedding vector of f_i. The attention layer, as shown in Fig 3, is formally defined by

$$\mathbf{u}_i^{h+1} = \phi_i^h \mathbf{u}_i^h, \tag{10}$$

$$\phi_i^h = \frac{exp(z_i^h)}{\sum_{j=1}^{|F_d|} exp(z_j^h)}, \tag{11}$$

$$z_i^h = \mathbf{u}_i^h \cdot M^h \cdot \bar{\mathbf{u}}^{h\top}, \tag{12}$$

$$\bar{\mathbf{u}}^h = \frac{1}{|F_d|} \sum_{i=1}^{|F_d|} \mathbf{u}_i^h, \tag{13}$$

where F_d denotes the feature set of a pair d and $M^h \in \mathbb{R}^{m \times m}$ denotes the hidden parameter matrix of layer h.

In each layer, We compute the means of $\mathbf{u}_i^h$, denoted by $\bar{\mathbf{u}}^h$, to capture the global context of an instance d. The matrix of M^h maps each feature embedding $\mathbf{u}_i^h$ to its global context $\bar{\mathbf{u}}^h$. For batch processing with a set of instances, the mask layer filters the features that do not occur in the instances. The model repeats the attention layer H times to maximally capture complex correlations between features. Finally, the model outputs the attention weight ϕ_i through a Softmax layer to re-weight feature influence. Specifically, it coverts the output vector $\mathbf{u}_i^H$ into a scalar output, denoted as $\phi_{f_i}(d)$, by a linear transformation as follows

$$\phi_{f_i}(d) = \mathbf{w}_{f_i} \cdot \mathbf{u}_i^{H\top} + b_{f_i}, \tag{14}$$

where $\mathbf{w}_{f_i} \in \mathbf{w}$ denotes a weight vector with the size of m and $b_{f_i} \in \mathbf{b}$ denotes a scalar bias.

Similar to the original GML, the process of attention-enhanced gradual inference essentially learns the parameter values, denoted by $\theta = (\alpha, \tau, M^0, \cdots, M^H, \mathbf{w}, \mathbf{b})$, such that the inferred results maximally match the evidence observations on labeled instances.

Empirical Evaluation

Our empirical study has been conducted on three real benchmark datasets, which are described as follows:

- DBLP-ACM[1] (denoted by DA): ER needs to match the publication entities between DBLP and ACM. After blocking, the workload consists of 6402 pairs, in which 2207 pairs are equivalent and the remaining 4195 ones are inequivalent.
- Abt-Buy[2] (denoted by AB): ER needs to match the product entities between Abt.com and Buy.com. After blocking, the workload consists of 8924 pairs, in which 774 pairs are

[1] available at https://dbs.uni-leipzig.de/file/DBLP-ACM.zip
[2] available at https://dbs.uni-leipzig.de/file/Abt-Buy.zip

equivalent and the remaining 8150 ones are inequivalent.

- Songs[3] (denoted by SG): ER needs to match the song records within the same table. After blocking, the workload consists of 8312 pairs, in which 5211 pairs are equivalent and the remaining 3101 ones are unmatched.

We have compared AGML with both unsupervised alternatives and supervised models. The unsupervised alternatives include

- Unsupervised Clustering (denoted by UC). It maps the record pairs into points in a multi-dimensional feature space and then uses the k-means technique to cluster them into distinct classes based on distance measurement.
- Unsupervised Rule-based (denoted by UR). It reasons about pair equivalence based on the predefined rules, which are specified in terms of record similarity. For fair comparison, our implementation estimates the proportion of equivalent and inequivalent instances based on the UC results, and then label the instances based on record similarity.
- Original Gradual Machine Learning (denoted by GML). Assuming that all features are independent, the original GML solution gradually labels the instances through iterative factor graph inference.

The supervised models include

- Support Vector Machine (denoted by SVM) [11]. Mapping record pairs to points in a multi-dimensional feature space, it first fits an optimal SVM classifier on labeled training data and then uses the trained model to label the pairs in test data.
- DeepMatcher (denoted by DM) [1]. It is the classical supervised DNN solution for ER.
- Ditto [2]. It is the state-of-the-art supervised DNN for ER based on pre-trained transformer-based language models. Ditto allows domain knowledge to be injected by highlighting the important pieces of input information that may be of interest to make labeling decisions.

As in [4], AGML uses pair similarity as the machine metric to identify easy instances. For fair comparison, given a percentage of easy instances (e.g. 30%), AGML first uses the result of unsupervised clustering (UC) to estimate the percentages of equivalent and inequivalent instances in a workload, and then proportionally identify the easy matching and unmatching instances by record similarity. Our implementation of AGML and the test datasets have also been made open-source available at the website[4].

Comparative Evaluation

In the comparative evaluation, we set the number of AGML's attention layers at 6 and its embedding size at 256. As shown in the subsection of Sensitivity Evaluation, the performance of AGML is to a large extent insensitive to the number of attention layers, and its performance is stable provided that embedding size is large enough. For fair comparison, we set other parameters of AGML to the same values as the original GML [4]. Specifically, the ratio of easy instances for both GML and AGML is set at 30%.

The detailed results are presented in Table 1(a), in which the best results of unsupervised and supervised approaches measured by F-1 on each dataset have been highlighted. The reported results of GML, AGML and supervised approaches are averages over ten runs. For the supervised approaches, we report their performance provided with different portions of training data. For instance, for SVM, "30%" means that 30% of a dataset are used for training; for DNN, "30%(25%:5%)" means that 25% of a dataset are used for model training and 5% are used for validation.

It can be observed that AGML performs considerably better than the unsupervised alternatives of UR, UC and GML. Specifically, AGML beats UR and UC by considerable margins on all the test datasets. Their performance differences in terms of F-1 are larger than 8% on AB and SG, and the margins are close to 4% on DA. AGML also consistently performs better than GML by the margins of 2.5%, 3.2% and 2.6% on DA, AB and SG respectively. Due to the inherent challenge of ER, these improvements are indeed considerable.

It can also be observed that even with the

[3] available at http://pages.cs.wisc.edu/~anhai/data/falcon_data/songs

[4] https://chenbenben.org/agml.html

Table 1: Evaluation results on different datasets

(a) Comparative Evaluation

	AGML			GML			UR			UC		
	recall	precision	F1	recall	precision	F1	recall	precision	F1	recall	precision	F1
DA	0.906	0.998	**0.941**	0.849	0.994	0.916	0.963	0.842	0.901	0.965	0.857	0.908
AB	0.691	0.559	**0.618**	0.623	0.554	0.586	0.773	0.300	0.432	0.800	0.311	0.448
SG	0.988	0.981	**0.984**	0.925	0.994	0.958	0.665	0.966	0.788	0.855	0.589	0.697

	SVM								
	10%			20%			30%		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
DA	0.881	0.993	0.937	0.889	0.990	0.937	0.952	0.989	0.970
AB	0.418	0.771	0.527	0.440	0.659	0.528	0.423	0.700	0.528
SG	0.955	0.999	0.976	0.957	0.999	0.977	0.964	0.998	0.981

	DM								
	10%(5%:5%)			20%(15%:5%)			30%(25%:5%)		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
DA	0.942	0.978	0.960	0.961	0.978	0.969	0.959	0.982	0.971
AB	0.043	0.254	0.074	0.441	0.601	0.509	0.444	0.707	0.546
SG	0.980	0.975	0.977	0.987	0.987	0.987	0.993	0.990	0.991

	Ditto								
	10%(5%:5%)			20%(15%:5%)			30%(25%:5%)		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
DA	0.902	0.950	0.925	0.985	0.965	0.975	0.988	0.976	**0.982**
AB	0.649	0.306	0.416	0.858	0.597	0.704	0.841	0.817	**0.829**
SG	0.962	0.942	0.951	0.965	0.992	0.978	0.992	0.993	**0.992**

(b) Sensitivity Evaluation w.r.t Attention Layers

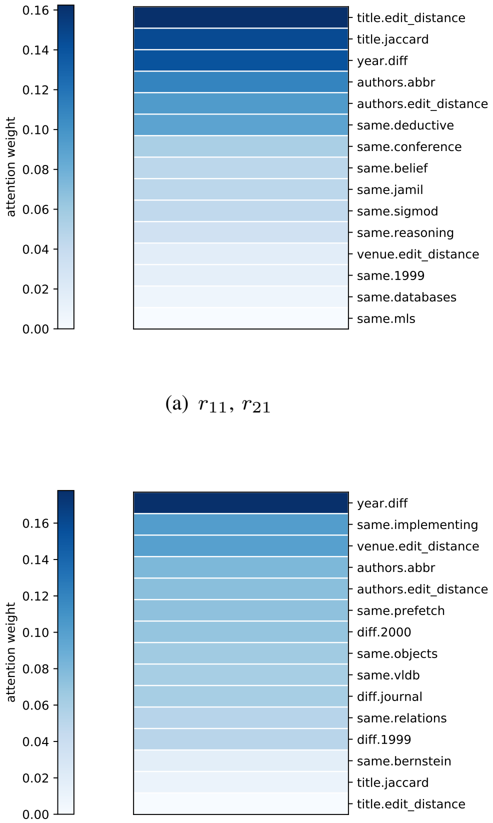
F-1	2	4	6
DA	0.941	0.942	0.941
AB	0.61	0.614	0.618
SG	0.976	0.985	0.984

(c) Sensitivity Evaluation w.r.t embedding size

F-1	64	128	256
DA	0.877	0.91	0.941
AB	0.573	0.619	0.618
SG	0.752	0.714	0.984

proportion of training data set at 30%, the performance of AGML is highly competitive with SVM and DM. For instance, with the proportion of training data set at 30%, AGML beats DM on AB while achieving competitive performance on DA and SG. It is noteworthy that on AB, AGML beats SVM and DM by considerable margins even with the 30% ratio of training data. The AB data contain only two important attributes for equivalence reasoning, product name and product description. Product names are usually very short, while the attribute of product description can contain short or long sentences. Such characteristics make the SVM and DM hard to accurately match records. Compared with Ditto, AGML achieves competitive performance when the proportion of training data is low (e.g. 10% and 20%). If provided with more training data (e.g. 30%), Ditto beats AGML in performance. However, it is worthy to point out that the efficacy of Ditto depends on sufficient labeled training data while no such data is supposed to be available for AGML.

To better understand the advantage of AGML over GML, we visualize the attention weights of



(a) r_{11}, r_{21}

(b) r_{12}, r_{22}

Figure 4: Weight Visualization.

the features in the running examples in Figure 4, in which color depth indicates weight value. It can be observed that on the pair of $< r_{11}, r_{21} >$, attention is focused on the features of similarities in title and author name, while on $< r_{11}, r_{21} >$, *year.diff* is correctly attentioned to be as the most decisive feature.

Sensitivity Evaluation

The results w.r.t the number of attention layers are presented in Table 1(b). Our experiment varied the number of attention layers from 2 to 6. It can be observed that the performance of AGML is very robust w.r.t the number of attention layers. With only 2 attention layers, AGML can achieve very competitive performance on all the test datasets.

The results w.r.t the embedding size are presented in Table 1(c). Our experiment varied the embedding size from 64 to 256. It can be observed that the performance of AGML is sensitive to the embedding size. When the embeddings size is 256, the performance of AGML is competitive and stable on all the 3 datasets. However, the performance of AGML generally deteriorates on all the three datasets when embedding size decreases. Its performance on SG decreases more dramatically because SG contains a larger number of tokens thus more features. As a result, SG requires a larger embedding space to learn a fine attention model. This observation should not come as a surprise, since it has been widely recognized that the capability of an attention space to capture feature correlation to a large extent depends on its space size.

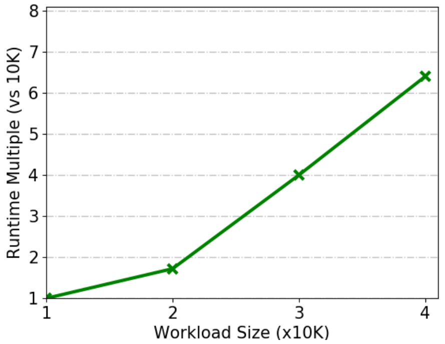


Figure 5: Scalability Evaluation.

Scalability Evaluation

For scalability evaluation, we generate different-sized DA workloads (between 10000 to 40000) based on the DBLP and ACM corpus. The results are presented in Figure 5, in which the x-axis denotes workload size and the y-axis denotes the cost multiple with the runtime spent on the workload of $10k$ as the baseline. It can be observed that the total consumed time increases nearly linearly with workload size. For a fixed batch size, the training time only increases with data size. Therefore, the attention-enhanced inference approach scales well with workload size.

Conclusion

In this paper, we have proposed a new attention-enhanced inference approach for gradual machine learning. Our extensive experiments have validated its efficacy. We have observed that pre-trained language models (e.g. BERT) are effective at improving the performance of deep

ER. While our current work generated feature representations based on feature co-occurrence, it is interesting in future to investigate how to improve feature representation by fusing the influence of feature co-occurrence and pre-trained language models.

REFERENCES

1. S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, 2018, pp. 19–34. [Online]. Available: https://doi.org/10.1145/3183713.3196926
2. Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *Proceedings of the VLDB Endowment*, vol. 14, no. 1, p. 5060, Sep 2020. [Online]. Available: http://dx.doi.org/10.14778/3421424.3421431
3. B. Hou, Q. Chen, J. Shen, X. Liu, P. Zhong, Y. Wang, Z. Chen, and Z. Li, "Gradual machine learning for entity resolution," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2019, pp. 3526–3530. [Online]. Available: https://doi.org/10.1145/3308558.3314121
4. B. Hou, Q. Chen, Y. Wang, Y. Nafa, and Z. Li, "Gradual machine learning for entity resolution," *IEEE Transactions on Knowledge and Data Engineering*, 2020. [Online]. Available: https://doi.org/10.1109/TKDE.2020.3006142
5. R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020. [Online]. Available: https://doi.org/10.1145/3406095
6. P. Christen, *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, ser. Data-Centric Systems and Applications. Springer, 2012. [Online]. Available: https://doi.org/10.1007/978-3-642-31164-2
7. Y. Wang, Q. Chen, J. Shen, B. Hou, M. Ahmed, and Z. Li, "Aspect-level sentiment analysis based on gradual machine learning," *Knowledge-Based Systems*, vol. 212, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.106509
8. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380
9. M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, 2010, pp. 1189–1197. [Online]. Available: http://papers.nips.cc/paper/3923-self-paced-learning-for-latent-variable-models
10. N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognition*, vol. 106, p. 107448, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132032030251X
11. P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 2008, pp. 151–159. [Online]. Available: https://doi.org/10.1145/1401890.1401913

Ping Zhong Ping Zhong is a Ph.D student in School of Computer Science in Northwestern Polytechnical University. His research interests include data quality, knowledge-based systems and artificial intelligence.

Zhanhuai Li Zhanhuai Li is a professor in School of Computer Science in Northwestern Polytechnical University. His research interests include data storage and management.

Qun Chen Qun Chen is a professor in School of Computer Science in Northwestern Polytechnical University. His current research interests include gradual machine learning and risk analysis for AI. He is the corresponding author of this article.

Boyi Hou Boyi Hou is a Ph.D student in School of Computer Science in Northwestern Polytechnical University. His research interests include data quality and artificial intelligence.