

State of the Art Instance Matching Methods for Knowledge Graphs^{*}

Alex Boyko^{1,2}[0000–0003–3592–4986], Siamak Farshidi²[0000–0001–6139–921X], and
Zhiming Zhao²[0000–0002–6717–9418]

¹ Vrije Universiteit Amsterdam, The Netherlands
`o.y.boyko@student.vu.nl`

² Universiteit van Amsterdam, The Netherlands
`{s.farshidi, z.zhao}@uva.nl`

Abstract. Instance matching identifies and matches instances in different data sources that refer to the same real-world entity. Many instance matching approaches have been proposed by the recent studies in the field. The goal of this literature review is to give an overview of the most recent studies that address instance matching, as well as to answer some of the research questions that are formulated based on the current research trends in this field. In this study we investigate and summarize seven state of the art instance matching approaches. We also investigate recent studies that touch upon active learning and privacy preservation. We conclude this study with an overview of the main findings presented in a diagram that depicts different pipelines used by the state of the art approaches.

Keywords: instance matching · asset discovery · entity alignment · IT asset linking · active learning · privacy-preserving algorithms

1 Introduction

Instance matching identifies and matches instances in different data sources that refer to the same real-world entity [30]. The instances that need to be matched usually come from one or multiple data sources [30]. In scientific literature, instance matching is also referred to as entity alignment [30], entity resolution [20], record linkage [11], data matching [6], entity reconciliation [8] and more. Typically, instance matching is used in databases tasks, such as deduplication, data merges, data processing, development and operations of different IT environments [6].

Matching instances generally yields three types of results: exact match, near match, and no match [27]. The threshold for each category should be tuned based on the underlying data[27]. Usually, for instance matching we are interested in the exact match pairs, however near-match results can also be useful, for example

^{*} Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

when applied in the context of recommendations [24]. No match pairs are also useful for training and tuning a classification model that can be used to perform instance matching [27].

The goal behind this research is to get a clear overview of the state of the art instance matching approaches and later apply them on real world data from ING Bank Netherlands. ING provided us with a number of real-world datasets that contain information about the IT assets used within ING. They have also provided us with a number applications for the instance matching implementation. These are:

1. Find a possible cause of an error.
2. Find and inform relevant stakeholders about an error.
3. Decommission an IT asset.
4. Give statistical insights about the the IT assets.

The contributions of this study can be summarized in two parts. (1) an example of using different keywords to explore state of the art approaches for instance matching, (2) a pipeline generalization of each studied state of the art approach, which can be used by other researchers as the basis for the research related to instance matching.

The remainder of this paper is organized as follows. Section 2 discusses the research approach. Section 3 presents the key findings of this research. Section 4 introduces the ING use case in more detail. Section 5 consists of the discussion of the main findings and threads to different types of validity. Section 6 discusses the recent work that has been done in this area. Section 7 concludes this paper by summarizing the main findings and giving directions for the future research.

2 Research Approach

In this section, we elaborate on the research methods and formulate three research questions that address state of the art instance matching. To answer these questions we conduct a literature study that provides a better view on the current state of instance matching and future directions.

2.1 Problem definition

With so many synonyms of instance matching, it can be difficult to get a clear overview of all recent approaches that have been published under different names, e.g. entity alignment. These synonyms are most likely aimed at solving the same kind of problem, while using different keywords for a search will retrieve a completely different list of results of the instance matching methods. Not being aware of different names of the same task can limit the choices of the researcher and possibly lead to a suboptimal method being implemented.

Moreover, there has been an increasing number of instance matching solutions that have been proposed in the past few years. Due to the large number and the diversity of the these methods, it can be difficult to choose the optimal approach to implement, especially without the sufficient knowledge.

2.2 Research Questions

The following research questions have been designed to gain in-depth insights about the recently developed instance matching approaches.

RQ1: *What are the state of the art approaches for instance matching across knowledge graph?*

Analyzing state of the art instance matching approaches helps us understand how different methods are used to achieve high performance, in terms of precision, recall and accuracy. Identifying the state of the art is not an easy task, since the performance of each instance matching approach highly depends on the usage of different types of information, for example during training or validation of a learning model.

RQ2: *How can active learning be used for instance matching?*

The term Active Learning is generally used to refer to a learning problem or a system where the learner has an important role in determining on what data the machine learning model will be trained [7]. An important feature of every instance matching method is its ability to retrieve and rank the similarity results accurately. However, setting too high of a threshold of accuracy can result in fewer results retrieved and a negative impact on user experience [2]. Sometimes there is an option to include human experts into the instance matching lifecycle, either during training, to help train the model when data is unclear, or during the testing to give the feedback to the model and help it to improve its future performance. In the ING case, we will assume that the experts are the people who create, use and/or maintain instances that are stored in the data sources.

RQ3: *How can data privacy be preserved during instance matching?*

In a non-privacy-preserving context, instance matching can be based on a shared personal identifiers, such as name, address, gender, and date of birth, which are effectively used as weak IDs [13]. However, in the privacy preserving scenario weak identifiers are considered private to each party. There is an increasing trend of concerns about privacy of the data that is being collected and shared by companies or online websites [9]. These concerns are related to both the consumer, whose information is often used or traded with little consent, and the collector, who is often responsible for protecting the collected data.

2.3 Research Method

Literature study consists of a manual search and automatic search. For instance matching manual search aims at looking into top tier journals and conferences that address knowledge and data related processes. The most prominent journals and conferences were chosen based on the (1) area of expertise and (2) its ranking on CORE Rankings Portal (A* or A).

The automatic search consists of multiple search queries that contain keywords related to the topic being researched and clauses (AND, OR, NOT, etc) that connect these keywords. The goal of these search query is to capture the information related to the research questions defined above.

Figure 1 shows the phases of the search protocol used for this literature study. The numbers between the parentheses denote a number of studies filtered after each phase of the search process.

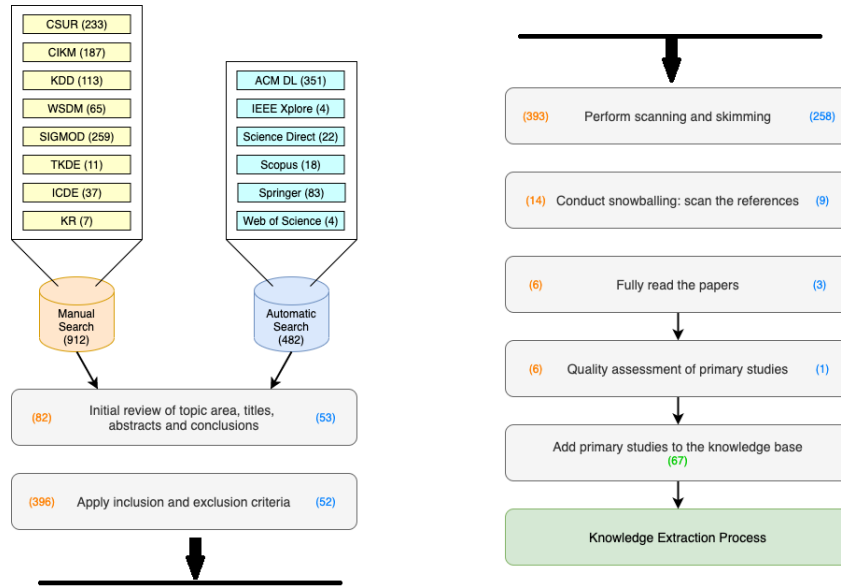


Fig. 1. Phases of the search process and the number of primary studies in each phase of the literature review.

3 Findings

The trend of instance matching has increased in the past few years. The difference in the number of papers related to instance matching being published with seventeen hundred papers being published in the year 2000 and almost ten thousand papers in the year 2020.

The main similarity of every state-of-the-art approach studied here is the use of the embeddings to represent the data in the form of vectors. The learned embeddings are then used to train a machine learning model. Ideally, the goal of the embeddings is to capture some of the semantics of the input data by placing semantically similar inputs close together in the vector space. Different approaches described in this section learn different kinds of embeddings (relation, entity, position, etc.) by using different methods (GCN, GNN, pre-trained, etc.). The are

many differences between the approaches, where the main ones are the kind of information is used to perform instance matching (relations, neighbors, structure, external ontology, etc.) and what kind of machine learning approaches are used (supervised learning, unsupervised, self-supervised, reinforcement, deep). This section gives a short summary of each of the state-of-the-art approaches identified in this study.

3.1 State of the Art Instance Matching

Some learning models can be trained by using the semantic information of the data, which enhances their ability to identify patterns and find similarities between pairs of entities. One of such approaches is proposed by Zhu et al. [30], which implements the relation-aware neighborhood matching (RNM) model that explores useful information from the connected relations. By giving two knowledge graphs (KGs) and a set of seed alignments of entities as input, it jointly learns the embeddings of entities and relations using graph convolutional networks (GCNs) with a TransE-like [4] regularizer. After that, it iteratively aligns the entities and relations in a semi-supervised manner. Each iteration utilizes the graph structure information to determine new matching pairs of entities and relations by combining relation-aware and entity-aware neighborhood matching modules.

Another way to learn graph embeddings is to develop a graph neural network (GNN) to rank soft correspondences between nodes. This approach is adopted in the study of Fey et al. [10], where they propose a two-stage neural architecture for learning and refining structural correspondences between graphs. After computing GNN-based localized node embeddings, they employ synchronous message passing networks to iteratively re-rank the soft correspondences to reach a matching consensus in local neighborhoods between graphs. They show that their underlying message passing scheme computes a well-founded measure of consensus for corresponding neighborhoods, which is then used to guide the iterative re-ranking process.

Many instance matching approaches are designed with specific applications in mind. For example, focused on applying instance matching for decision-making Zeng et al. [28] present a model based on reinforcement learning (CEAFF). Using the reinforcement learning framework, they devise the coherence and exclusiveness constraints to characterize the interdependence and restrict collective alignment. Additionally, to feed more precise inputs to the reinforcement learning model, they employ representative features to capture different aspects of the similarity between entities in heterogeneous KGs, which are integrated by an adaptive feature fusion strategy.

Another embedding-based model (COTSAE) is proposed by Yang et al. [26], where it is proposed to combine the structure and attribute information of entities by co-training two embedding learning components. They also propose a joint attention method for the model to learn the attentions of attribute types and values cooperatively. Tang et al. [25] present an interaction model that focuses on leveraging the semantic information. Instead of aggregating neighbors,

they compute the interactions between neighbors that can capture fine-grained neighbors' matches while also modeling the interactions of attributes.

Existing supervised methods for instance matching require a training dataset to be provided to them, and they focus on pulling each pair of positively labeled entities close to each other. However, the learning of instance matching can possibly benefit more from pushing sampled (unlabeled) negatives far away than pulling positively aligned pairs close, according to Liu et al. [17]. Their study presents yet another model (SelfKG), a self-learning model that leverages this discovery to design a contrastive learning strategy across two KGs. A similar solution proposed by Zeng et al. [27] offers an unsupervised framework that performs entity alignment (UEA) in the open world. Specifically, it first mines useful features from the entity information present within the KGs, such as the attributes, descriptions, and classes. Then, it devises an unmatchable entity prediction module to filter out unmatchable entities and produce preliminary alignment results. These preliminary results are regarded as pseudo-labeled data and forwarded to the progressive learning framework to generate structural representations, which are integrated with the entity information to provide a more comprehensive view for alignment. Finally, the progressive learning framework gradually improves the quality of structural embeddings and enhances the alignment performance by enriching the pseudo-labeled data with alignment results from the previous round. This solution automatically generates labeled data and succeeds at effectively filtering out unmatchable entities.

From the study of Pei et al. [21] about graph neural networks for cross-lingual instance matching between knowledge graphs, it is clear that their model can be improved by finding a reliable way to reduce human effort in creating a set of trusted entity pairs as the positive samples. Moreover, their graph neural network approach misrecognizes some labeled entity pairs as noisy pairs and some noisy entity pairs as real pairs, which means it is possible to study how to reduce the effect of misrecognized entity pairs on the performance of instance matching. Finally, it is also possible to expand the exploration and discussion about noise in the attribute of entities and knowledge graphs.

With regard to reinforcement learning for instance matching [28] it is possible to research more advanced feature encoders that can better exploit available features for alignment. In addition, it is worth exploring collective alignment algorithms that can further mitigate the error propagation caused by the wrong matches.

3.2 Active Learning

While datasets used for benchmarking usually contain the ground truth and training data, most real-world datasets do not. In reality, creating alignments that can be used to train a machine learning model is time consuming and often requires human annotators. [1] This is why active learning for instance matching has received a lot of attention in the recent years. Some papers propose approaches that focus on the labeling stage [1] in the instance matching pipeline and some focus more on the matching stage [14]. Most recent studies implement

deep active learning, however one study describes an approach that uses pre-trained language models. This section focuses on some of the most prominent approaches in the fields of active learning and instance matching.

A study done by Jain et al. [14] proposes an instance matching model (DIAL), which is based on scalable active learning that jointly learns embeddings to maximize recall for blocking and accuracy for matching blocked pairs. They claim that passive learning methods on tasks like instance matching require large amounts of labeled data to yield useful models, while active learning is a promising approach in low resource settings.

On the contrary, Berrendorf et al. [1] claim that by testing different active and passive learning strategies, they could conclude that passive learning approaches, which can be efficiently precomputed and deployed, achieve comparable performance to the active learning strategies.

Many other instance matching approaches have been reported to use active learning within its pipeline. Bogatu et al. [3] use active learning to reduce the cost of labeling training data through an approach that builds on the properties conferred by the use of deep autoencoders. Li et al. [16] integrate active learning into their method of entity similarity calculation for knowledge graphs. Kasai et al. [15] implement a low-resource deep instance matching model with transfer and active Learning. Loster et al. [18] use active learning to construct deep Siamese neural networks, capable of learning a similarity measure that is tailored to the characteristics of a particular dataset. Nafa et al. [19] propose an active deep learning model of risk sampling for instance matching. Risk sampling aims at sampling challenging examples for the classifier and produce better representation learning by looking farther than the low confidence regions in the current representation space. Finally, Qian et al. [23] propose an active learning model for large-scale instance matching, which learns, multiple rules each having significant coverage of the space of entity matches. An interesting direction for the research in active learning is to find a good way to integrate transfer learning, which will allow to reuse the trained models to reduce the amount of required labels for the future models.

3.3 Privacy-preserving Instance Matching

The approaches described in the scientific literature have different applications. For example, Peng et al. [22] focus on the data that comes from multiple sources and how it affects the privacy of that data. Additionally, Zhang et al. [29] present the dangers to privacy during the semantic social network instance matching, which deals with user-sensitive data. The rest of this section describes in more detail how privacy is treated and what are the dangers to privacy during the semantic instance matching.

The decentralized environment can pose some serious privacy related issues, where some nodes in the system do have access to certain private information and some do not. This is why the federated knowledge graph embedding framework proposed by Peng et al. [22] also addresses privacy preservation in its implementation. Given two KGs with aligned entities and relations, FKGE exploits

a Generative Adversarial Net (GAN) [12] structure to unify the embeddings of aligned entities and relations. After GAN training, the synthesized embeddings are able to learn features from both KGs and therefore can replace original embeddings as refined and unified embeddings. It is sufficient for GAN training that only the generated outputs and gradients are transmitted without revealing raw data. However, even for generated embeddings, there are still privacy concerns for reconstruction attacks. It is possible that neural models may memorize inputs and reconstruct inputs from corresponding outputs [5]. To further address the privacy issue, they introduce differential privacy to privatize generated embeddings. Differential privacy provides a strong guarantee for protecting any single embedding in the generator outputs since inclusion and exclusion of a particular embedding will not affect the outcome distribution too much.

4 ING IT Asset Example

ING is an example of an organization with a complex IT landscape that is being maintained by developers all over the world. ING has to keep track and carefully manage the data related to its IT assets to be able to orchestrate different components of the system. The term IT asset is a broad term, which is used to referring to a business application, micro-service, API, network server, virtual machine, and more.

As part of this project ING has provided us the access to the internal data that contains instances that should be match. This data is stored in Neo4j, a database management system that stores the data as a graph. In such database IT assets are represented as nodes and the relationships between assets are represented as edges. Both nodes and edges can have any number of attributes stored. Since the data is stored in a graph format, it makes it applicable for the knowledge based instance matching state of the art methods (RQ1). IT asset data is actively used and maintained by many teams within ING. This provides an opportunity to learn useful insights about the data to obtain rich training/validation datasets for the instance matching model (RQ2). In general, IT asset data may contain sensitive information, for example:

1. information about servers (IP address).
2. information about software and IT products (incidents, security, dependencies)

Addressing privacy-preservation during instance matching can help with avoiding any data leakage or misuse of private data in the future (RQ3).

5 Discussion

Figure 2 depicts a swimlane-like diagram representing different workflow pipelines used by different state of the art instance matching approaches. This diagram should serve as a basis for the researchers that are interested in instance matching or related fields.

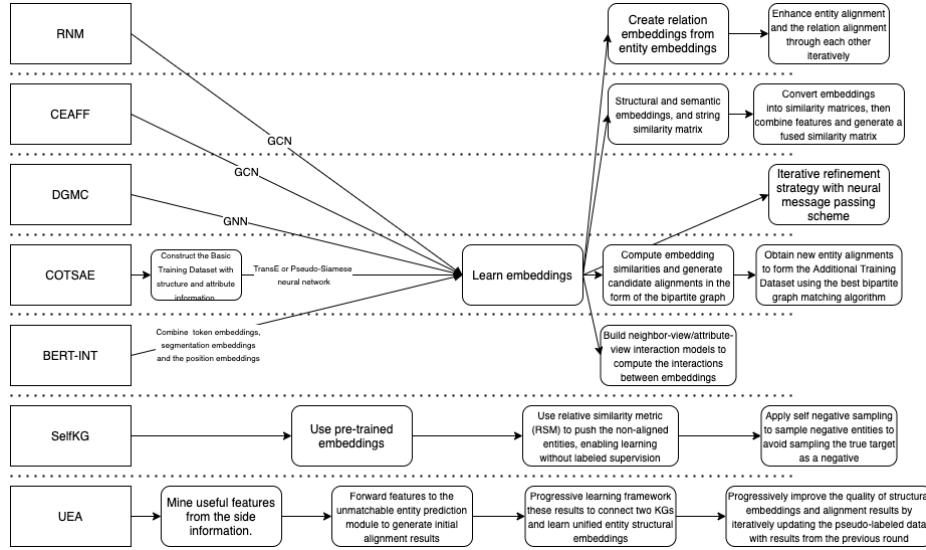


Fig. 2. A swimlane-like diagram that depicts workflow pipelines of the state of the art instance matching approaches.

5.1 Threads to Validity

External validity In this study the main external threat to validity may come from the primary studies not being representative of the whole research on instance matching. This is expected, because instance matching has many alternative names, thus it is difficult to find all the recent studies related to it.

Internal validity The threats to internal validity target the design and execution of the literature review. To prevent possible design errors, the systematic literature review protocol has been followed. In addition, to be able to select the best paper on instance matching, the inclusion/exclusion criteria was designed and.

Construct validity The threats to construct validity target the alignment between the research question and the measurements of the review. We evaluate the effectiveness of our search query with a list of pilot studies that has proven to answer the research query based on the additional checks and validations of the experts in the field.

Conclusion validity The threats to conclusion validity come from the the data extraction process not being reproducible. For this study each data extraction step has been recorded in detail and we could reproduce all the steps successfully and get the same data.

6 Related Work

In their study Papadakis et al.[20] provide a broad literature coverage which serves as a basis for designing JedAI, an Entity Resolution tool that can also be used as a library of the main established techniques in the literature. In their research Papadakis et al. cover both application in academic and commercial scenarios to produce an application that incorporates methods that support four different end-to-end Entity Resolution workflows. This application integrates the following state-of-the-art ER techniques: budget-agnostic based on batch, schema-agnostic based on blocks and schema-based relying on similarity joins. As part of JedAI Papadakis et al. also offer the ability to build and benchmark millions of ER pipelines, applications to structured and semi-structured data, running on stand-alone and cluster computers through parallel implementation and adaption to budget-aware ER.

Reynoso and Divan[24] present a systematic mapping study, that addresses semantic similarity of entities and the reuse of the knowledge and previous experiences when a new entity has either none or little evidence, but is similar to another entity. Thus, when designing a system focused on decision making, the reuse of the knowledge and previous experiences plays an important role, especially for recommendations provided to the user. The study concludes that the semantic similarity applied to entities under monitoring in the measurement and evaluation projects is a challenge. In the end, the authors suggest a comparative analysis among the strategies, functionalities, and algorithms related to the semantic similarity, however they do not mention the importance of certain semantic matching properties, like privacy preservation, where the reuse of the previous knowledge should not be able to reveal any sensitive information.

7 Conclusion

Instance matching is a trending research field with many recent studies published in highly ranked journals and conferences. Instance matching aims at linking data that comes from different data sources based on their similarity. This study conducts a literature review to give an overview of the most recent studies in the field of instance matching, as well as answer some of the research questions that are formulated based on the current research trends in the field. Three research question are considered when constructing search queries used to explore research studies published in some of the highly ranked journals and conferences and to retrieve 1,394 research studies. After going through the phases of the search and applying inclusion/exclusion criteria 67 primary studies have been selected to answer the research questions from this study.

For future directions it is possible to extend this study by researching other synonyms of instance matching, for example entity resolution or record linkage. Another direction for future work is to explore different application of instance matching, for example in Semantic Web, recommender systems, document alignment, applications in pharmaceutical and biomedical sciences and more.

References

1. Berrendorf, M., Faerman, E., Tresp, V.: Active learning for entity alignment. In: ECIR (2021)
2. Berrendorf, M., Faerman, E., Tresp, V.: Active learning for entity alignment. CoRR **abs/2001.08943** (2020), <https://arxiv.org/abs/2001.08943>
3. Bogatu, A., Paton, N., Douthwaite, M., Davie, S., Freitas, A.: Cost-effective variational active entity resolution. ArXiv **abs/2011.10406** (2020)
4. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS (2013)
5. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting training data from large language models. ArXiv **abs/2012.07805** (2020)
6. Christen, P.: Data Matching (01 2012). <https://doi.org/10.1007/978-3-642-31164-2>
7. Cohn, D.: Active Learning, pp. 9–14. Springer US, Boston, MA (2017). https://doi.org/10.1007/978-1-4899-7687-1_9, https://doi.org/10.1007/978-1-4899-7687-1_9
8. Enríquez, J., Domínguez-Mayo, F., Escalona, M., Ross, M., Staples, G.: Entity reconciliation in big data sources: A systematic mapping study. Expert Systems with Applications **80**, 14–27 (2017). <https://doi.org/10.1016/j.eswa.2017.03.010>, <https://www.sciencedirect.com/science/article/pii/S0957417417301550>
9. Esperança, P., Aslett, L.J.M., Holmes, C.C.: Encrypted accelerated least squares regression. In: AISTATS (2017)
10. Fey, M., Lenssen, J.E., Morris, C., Masci, J., Kriege, N.M.: Deep graph matching consensus. ArXiv **abs/2001.09621** (2020)
11. Gautam, B., Terrades, O.R., Pujades, J.M., Valls, M.: Knowledge graph based methods for record linkage. ArXiv **abs/2003.03136** (2020)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
13. Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., Thorne, B.: Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. ArXiv **abs/1711.10677** (2017)
14. Jain, A., Sarawagi, S., Sen, P.: Deep indexed active learning for matching heterogeneous entity representations. ArXiv **abs/2104.03986** (2021)
15. Kasai, J., Qian, K., Gurajada, S., Li, Y., Popa, L.: Low-resource deep entity resolution with transfer and active learning. ArXiv **abs/1906.08042** (2019)
16. Li, L., Zhang, Z., Zhang, S.: Knowledge graph entity similarity calculation under active learning. Complex. **2021**, 3522609:1–3522609:11 (2021)
17. Liu, X., Hong, H., Wang, X., Chen, Z., Kharlamov, E., Dong, Y., Tang, J.: A self-supervised method for entity alignment (2021)
18. Loster, M., Koumarelas, I.K., Naumann, F.: Knowledge transfer for entity resolution with siamese neural networks. Journal of Data and Information Quality (JDIQ) **13**, 1 – 25 (2021)
19. Nafa, Y., Chen, Q., Chen, Z., Lu, X., He, H., Duan, T., Li, Z.: Active deep learning on entity resolution by risk sampling. ArXiv **abs/2012.12960** (2020)
20. Papadakis, G., Mandilaras, G., Gagliardelli, L., Simonini, G., Thanos, E., Giannakopoulos, G., Bergamaschi, S., Palpanas, T., Koubarakis, M.: Three-dimensional entity resolution with jedai. Information Systems **93**, 101565 (2020). <https://doi.org/10.1016/j.is.2020.101565>, <https://www.sciencedirect.com/science/article/pii/S0306437920300570>

21. Pei, S., Yu, L., Yu, G., Zhang, X.: Rea: Robust cross-lingual entity alignment between knowledge graphs. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020)
22. Peng, H., Li, H., Song, Y., Zheng, V., Li, J.: Federated knowledge graphs embedding. *ArXiv* **abs/2105.07615** (2021)
23. Qian, K., Popa, L., Sen, P.: Active learning for large-scale entity resolution. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017)
24. Reynoso, M., Diván, M.: Assessment of semantic similarity in entities under monitoring: A systematic literature mapping. *Revista Facultad de Ingeniería Universidad de Antioquia* (05 2020). <https://doi.org/10.17533/udea.redin.20200476>
25. Tang, X., Zhang, J., Chen, B., Yang, Y., Chen, H., Li, C.: Bert-int: A bert-based interaction model for knowledge graph alignment. In: *IJCAI*. pp. 3174–3180 (2020)
26. Yang, K., Liu, S., Zhao, J., Wang, Y., Xie, B.: Cotsae: Co-training of structure and attribute embeddings for entity alignment. In: *AAAI* (2020)
27. Zeng, W., Zhao, X., Tang, J., Li, X., Luo, M., Zheng, Q.: Towards entity alignment in the open world: An unsupervised approach. In: *DASFAA* (2021)
28. Zeng, W., Zhao, X., Tang, J., Lin, X., Groth, P.T.: Reinforcement learning based collective entity alignment with adaptive features. *ArXiv* **abs/2101.01353** (2021)
29. Zhang, Y., Fu, J., Yang, C., Xiao, C.: A local expansion propagation algorithm for social link identification. *Knowledge and Information Systems* **60**, 545–568 (2018)
30. Zhu, Y., Liu, H., Wu, Z., Du, Y.: Relation-aware neighborhood matching model for entity alignment. In: *AAAI* (2021)