An Analysis of Analyses in Entity Alignment

Bryan Linares
bryan.linares01@student.csulb.edu
Ky-Thuyen Vugia
ky-thuyen.vugia01@student.csulb.edu

October 21, 2024

1 Iterative Geographic Entity Alignment with Cross-Attention Introduction

The paper Iterative Geographic Entity Alignment with Cross-Attention (IGEA) Dsouza, Yu, Windoffer, and Demidova (2023) covers a novel method for aligning knowledge graphs to geographic data. The importance of incorporating geographic data into usable semantic representations can lead to better location question answering and point of interest recommendations that draw on semantic knowledge graphs(KG). Other proposed systems based on RDF have been proposed but not adopted widely like LinkedGeoData(Stadler, Lehmann, Höffner, & Auer, 2012). Yet the most popular general purpose KGs like Wikidata and DBpedia have limited geographic data. A rich source of free geographic data on the web is Open Street Maps(OSM) but it is community-driven, leading to sparse and uneven entries and no clear interface to the general KGs. For instance, OSM uses tag pairs instead of the entities and classes of the general KGs. IGEA performs a method for tag-to-class and entity alignment iteratively, with candidate choosing mechanisms for efficiency and noise reduction.

The proposed end-to-end IGEA system has a few broad steps: first it performs Geographic Class Alignment through word embeddings to narrow the possible classes, then Generates Candidates using geographic distance, and the main Entity Alignment module performs the Entity Alignment using a layered Cross-Attention machine learning model before iteration. Results against other methods are shown to be promising.

2 Background

2.1 LSTM Model

The alignment module uses in part the Long Short Term Memory type of machine learning model, which is popular in text based applications within Natural Language Processing. An LSTM is a type of Recurrent Neural Network, which are characterized in general by having a ”memory” within their variables that allows data from the previous steps in a sequence to come forward to the current one. In the figure, an LSTM cell input has its current input, hidden input (h), and its context(c) which is the memory. The input and ”forget”(memory) gates decide the hidden output and memory of the cell that goes to the next. In IGEA, a Bi-directional LSTM is used that has the extra connection needed that allows processing of inputs forwards and backwards which provides a better representation of the sequence that can go both ways in the order. Such an arrangment has been shown to work well for named entity recognition(Chiu & Nichols, 2016).In IGEA, this helps align entities by finding elements of their representations in the input groups of entities that correspond to each other, as they are all grouped and concatenated together.

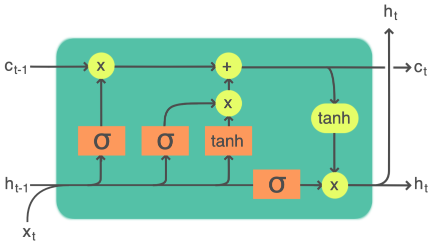


Figure 1: LSTM Cell

2.2 Cross-Attention Model

IGEA employs a layered Cross-Attention model for doing the ultimate entity alignment. The goal of IGEA is to align heterogeneous representations of the same semantic entity across OSM and KGs. An Attention model is a family of machine learning models that focus on selected parts of an input sequence. The model prioritizes and highlights the most relevant information in this way. A Cross-Attention model Lin et al. (2021) uses the Attention mechanism but also takes the most relevant parts of one input sequence while processing another. In IGEA the Bi-LSTM output serves as the input to the cross-attention layer. In figure [2] the current hidden states from the decoder are input to the weights calculation to be used as keys,queries, and values, the first two of which are put through a softmax to normalize the attention weights. The values are saved to the hidden states. These values put together, become a vector that summarizes the most important information found. In IGEA this helps learn the important tags or properties within an entity regardless of the other.

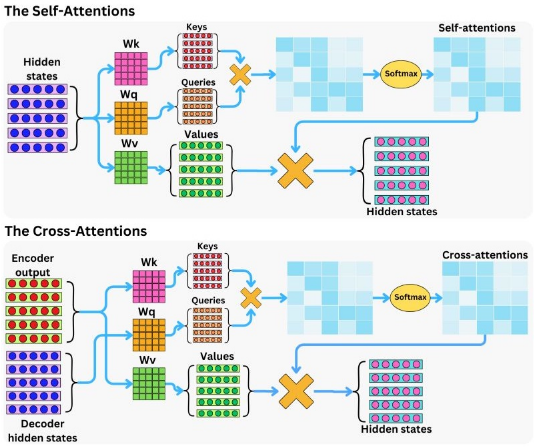


Figure 2: Self and Cross Attention Models

3 Method

3.1 Problem Statement

IGEA aims to perform the process of geographic entity alignment, interlinking them, connecting the general knowledge graphs to the community-driven data of OSM. Take the following example, in figure 3 the OSM tags for Berling are listed against the Wikidata triples on the right. The goal in simplest terms is that the OSM ”name” for Berlin will be linked to the Wikidata ”label” Berlin with a ”sameAs” link by Geographic Entity Alignment. The tag place=city in OSM is decided to be linked to the city(wd:Q515) class of Wikidata by Geographic Class Alignment. The geographic entity alignment is done through iterative learning of class and entity alignment.

(a) OSM tags.

Key	Value
name	Berlin
place	city
population	3769962
way	POINT(52.5183 13.4179)
capital	yes

(b) Wikidata triples. wd:Q64 identifies Berlin.

Subject	Predicate	Object
wd:Q64	rdfs:label (*label*)	Berlin
wd:Q64	wdt:P31 (*instance of*)	wd:Q515 (*city*)
wd:Q64	wdt:P1082 (*population*)	3677472
wd:Q64	wdt:P625 (*coordinate location*)	52°31'N, 13°23'E
wd:Q64	wdt:P1376 (*capital of*)	wd:Q183 (*Germany*)

Figure 3: Example for IGEA

3.2 IGEA Approach

There are several broad phases to IGEA as seen in overview [4]. First, known linked entities are gathered between OSM and KG from previous iterations. Using an approach NCA Dsouza, Tempelmeier, and Demidova (2021) based on word embeddings, the tags from OSM are aligned to the classes in the KG. These alignments are passed to a module with takes the tag-to-class alignments and checks their corresponding entities for Geographic data, and performs blocking (only certain candidates are used in the steps ahead) based on a threshold of geographic distance(in this case 2500m). The candidate pairs are passed into the main Cross-attention based entity alignment classifier. New matches are found and high confidence pairs are passed forward to new iterations.

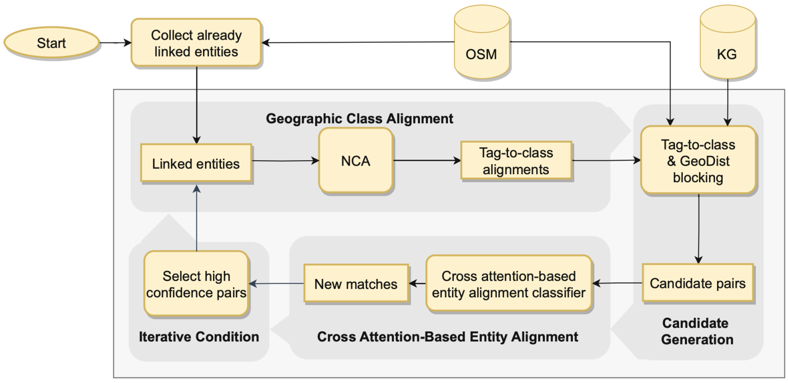

Start
Collect already linked entities
OSM
KG
Geographic Class Alignment
Linked entities
NCA
Tag-to-class alignments
Tag-to-class & GeoDist blocking
Select high confidence pairs
New matches
Cross attention-based entity alignment classifier
Candidate pairs
Iterative Condition
Cross Attention-Based Entity Alignment
Candidate Generation

Figure 4: Overview of IGEA

The Entity Alignment Module in 5 contains the machine learning classification model layers that are core to IGEA. In 5 the Entity Representation Module, the model input is prepare. For a given OSM node, all tags are concatenated together to make a sentence. For a KG entity, all predicates and objects are selected and concatenated into their own sentence. The average number of words of all candidates in the set is found, and the representation layer receives them to pass into the LSTM (pretrained word embeddings from fastText are used here). Then in the Cross-Attention Module the Bi-LSTM layer for the KG and OSM perform named entity recognition, since the model is meant to learn what comes after a particular key or property to help in the cross layer. All the hidden states are passed in. The cross-attention layer helps understand the important properties and tags for aligning the entities. Attention scores are built using the keys, values, and queries generated for each of the OSM and KG entries. The self-attention layer is used so the model can learn the important tags and properties of a given entity. Instead of the combination of outputs from the OSM and KG layers, only the single input from the OSM or KG is used, and the result is passed through a final LSTM layer. In the Classification Module the linked entities are used at the supervision for the classification. Each true pair is labeled one, and the remaining are labeled zero. This layer predicts whether a given pair is really a match or not, which is passed through some fully connected layers for a final score.

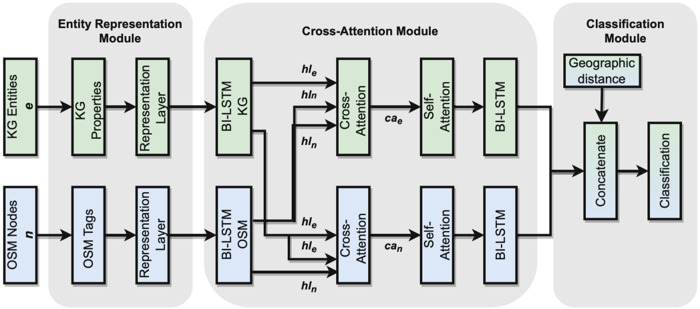
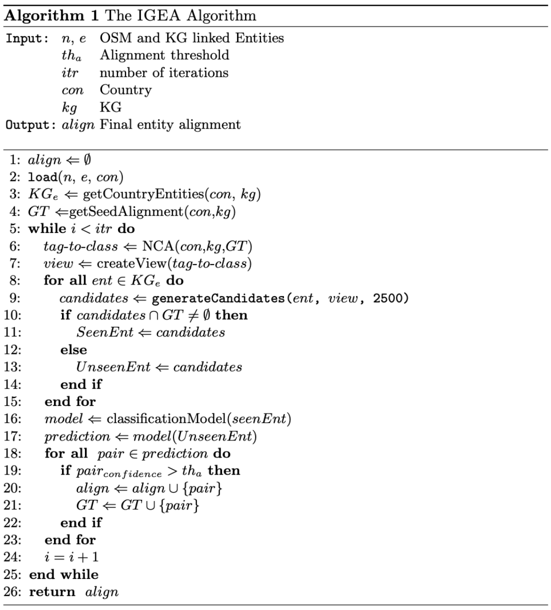


Figure 5: Entity Alignment Module

The Iterative IGEA approach is applied at the country level. For each country all entities with geo coordinates are taken from the KG. Once pairs of candidates are chosen, there is a threshold setting that can be adjusted to only select high confidence pairs to go through the process again.

3.3 IGEA Algorithm

The overall IGEA algorithm is contained in the figure [6] which serves as a reference for the code implementation and a summary of the above overview description.



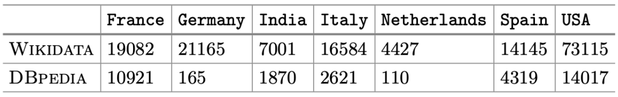
```
Algorithm 1 The IGEA Algorithm
Input:  $n$, $e$    OSM and KG linked Entities
        $th_a$      Alignment threshold
        $itr$       number of iterations
        $con$       Country
        $kg$        KG
Output: $align$ Final entity alignment

$align \Leftarrow \emptyset$
load($n$, $e$, $con$)
$KG_e \Leftarrow$ getCountryEntities($con$, $kg$)
$GT \Leftarrow$getSeedAlignment($con$,$kg$)
while $i < itr$ do
    $tag\text{-}to\text{-}class \Leftarrow$ NCA($con$,$kg$,$GT$)
    $view \Leftarrow$ createView($tag\text{-}to\text{-}class$)
    for all $ent \in KG_e$ do
        $candidates \Leftarrow$ generateCandidates($ent$, $view$, 2500)
        if $candidates \cap GT \neq \emptyset$ then
            $SeenEnt \Leftarrow candidates$
        else
            $UnseenEnt \Leftarrow candidates$
        end if
    end for
    $model \Leftarrow$ classificationModel($seenEnt$)
    $prediction \Leftarrow model(UnseenEnt)$
    for all $pair \in prediction$ do
        if $pair_{confidence} > th_a$ then
            $align \Leftarrow align \cup \{pair\}$
            $GT \Leftarrow GT \cup \{pair\}$
        end if
    end for
    $i = i + 1$
end while
return $align$
```

Figure 6: IGEA Algorithm

4 Results

The results evaluated considered OSM, Wikidata, and DBPedia datasets from various countries including France, Germany, and Spain. The OSM and KG data are loaded to a local database. As a ground truth, only country datasets with entities with more than 1500 geo-coordinate entries are chosen to have enough sufficient data to train the model.

4.1 Evaluation



	France	Germany	India	Italy	Netherlands	Spain	USA
Wikidata	19082	21165	7001	16584	4427	14145	73115
DBpedia	10921	165	1870	2621	110	4319	14017

Figure 7: Ground Truth amounts for Training and Eval

Evaluations were done against several baseline approaches to compare to. Some were word similarity based and others were also machine learning based. Here we will focus on the results against a closely related approach called OSM2KG, which is the best performing. OSM2KG uses a ML model that generates key-value embeddings Tempelmeier and Demidova (2021) based on the occurences of the tags including type and popularity but using a random forest classification model. Metrics use the F1 score which is the harmonic mean of both precision and recall. Precision is the ratio of all correctly identified pairs to all identified pairs. Recall is the fraction of all correct pairs to all pairs in the ground truth alignment.

In figure [8] the performance on the Wikidata KG is shown. IGEA is shown to be superior and performs consistently. It achieves up to an 18 percent improvement on the best performing baseline on the whole. The first and third iterations of IGEA are shown and they show improvement. The France and Spain datasets gain the most improvement. India seems to have the lowest available datasets and has the most limited performance increase. Figure [9] shows the effects of iteration of IGEA which gets to a maximum after only a few loops.

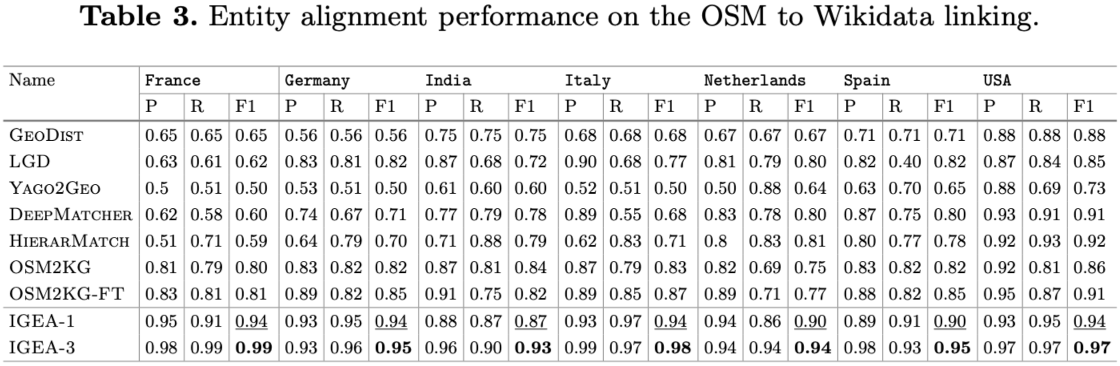
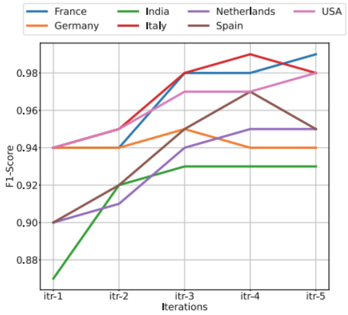


Table 3. Entity alignment performance on the OSM to Wikidata linking.

Name	France			Germany			India			Italy			Netherlands			Spain			USA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GeoDist	0.65	0.65	0.65	0.56	0.56	0.56	0.75	0.75	0.75	0.68	0.68	0.68	0.67	0.67	0.67	0.71	0.71	0.71	0.88	0.88	0.88
LGD	0.63	0.61	0.62	0.83	0.81	0.82	0.87	0.68	0.72	0.90	0.68	0.77	0.81	0.79	0.80	0.82	0.40	0.82	0.87	0.84	0.85
Yago2Geo	0.5	0.51	0.50	0.53	0.51	0.50	0.61	0.60	0.60	0.52	0.51	0.50	0.50	0.88	0.64	0.63	0.70	0.65	0.88	0.69	0.73
DeepMatcher	0.62	0.58	0.60	0.74	0.67	0.71	0.77	0.79	0.78	0.89	0.55	0.68	0.83	0.78	0.80	0.87	0.75	0.80	0.93	0.91	0.91
HierarMatch	0.51	0.71	0.59	0.64	0.79	0.70	0.71	0.88	0.79	0.62	0.83	0.71	0.8	0.83	0.81	0.80	0.77	0.78	0.92	0.93	0.92
OSM2KG	0.81	0.79	0.80	0.83	0.82	0.82	0.87	0.81	0.84	0.87	0.79	0.83	0.82	0.69	0.75	0.83	0.82	0.82	0.92	0.81	0.86
OSM2KG-FT	0.83	0.81	0.81	0.89	0.82	0.85	0.91	0.75	0.82	0.89	0.85	0.87	0.89	0.71	0.77	0.88	0.82	0.85	0.95	0.87	0.91
IGEA-1	0.95	0.91	0.94	0.93	0.95	0.94	0.88	0.87	0.87	0.93	0.97	0.94	0.94	0.86	0.90	0.89	0.91	0.90	0.93	0.95	0.94
IGEA-3	0.98	0.99	**0.99**	0.93	0.96	**0.95**	0.96	0.90	**0.93**	0.99	0.97	**0.98**	0.94	0.94	**0.94**	0.98	0.93	**0.95**	0.97	0.97	**0.97**

Figure 8: Entity Alignment performance on WD



(a) Wikidata

Figure 9: Iteration of IGEA on WD

Ablation tests were performed, where several of the novel parts of IGEA are removed and the results measured. The geographic distance removal leads to worse performance. This result highlights the importance of specific domain data in this case that depends on geographic correctness. Removal of the class-based blocking gets better recall but a drop in precision. This may mean that class-based blocking can effectively reduce amount of data to search and improve efficiency, but it can also introduce bad matches by incorrectly grouping entities based on bad results. Removing Cross-Attention decreases performance. This shows its importance in dealing with the heterogeneity, it allows the model to selectively focus on relevant parts of the entity descriptions (OSM tags and KG properties).

Name	France			India			Italy			Spain			USA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
w/o Cross-Attention	0.86	0.81	0.83	0.83	0.82	0.82	0.86	0.77	0.81	0.82	0.81	0.81	0.83	0.84	0.83
w/o Distance	0.85	0.89	0.86	0.81	0.87	0.82	0.81	0.83	0.82	0.79	0.86	0.82	0.82	0.87	0.84
w/o Class-Blocking	0.81	0.93	0.87	0.73	0.94	0.82	0.78	0.93	0.85	0.75	0.92	0.83	0.79	0.96	0.86
IGEA-3	0.95	0.99	**0.97**	0.96	0.97	**0.97**	0.95	0.98	**0.96**	0.96	0.95	**0.95**	0.99	0.97	**0.98**

Figure 10: Ablation Study

5 Conclusion and Future Work

5.1 Conclusion

IGEA results show that the novel Attention models are valuable in this context because they handle uneven and heterogeneous representations of data. They also cope with the missing data and sparsity in OSM by focusing on available data. The crucial elements are focused on and lead to more accurate results. IGEA is highly effective for geographic entity alignment. It consistently outperforms the baselines. The iterative nature adds to its gains in performance. The type of model is also crucial to getting the best results. IGEA is also capable of new links and fixing other wrong links on unseen data. Yet it can be weak based on the characteristics of its initial data.

5.2 Related and Future Work

The topics of Geographic Entity Alignment, ontology and schema alignment, and iterative learning remain very fruitful for research. IGEA contributes by aligning geographic data, and using deep learning to good result. Working with OSM and data without rich entries is very useful. Aligning the schema between the ontologies is usually done with structural and linguistic approaches but the deep learning approach here is potentially very powerful. In future work, some similar entries can be parsed more accurately, and working with non-English entries is a great point of effort. It could even be possible to incorporate external knowledge sources. IGEA is a strong foundation for advancing robust and scalable geographic entity matching and alignment.

6 Structural Bias in Knowledge Graphs Paper Introduction

The paper Structural Bias in Knowledge Graphs for the Entity Alignment Task focuses on determining whether the structure of current knowledge graphs (KGs) causes biases in the KG entity alignment (EA) task, and proposes a novel exploration-based sampling method (SUSIE) for improving performance of state-of-the-art KG EA methods and reducing their bias, specifically the bias regarding finding pairs of nodes in two input KGs that refer to the same real-world entity. These issues stem from inaccuracies from sampling methods such as Fairwalk, IDS, and div2vec as well as the EA methods applied afterwards such as RREA, OpenEA, MultiKE, and PARIS. EA tasks are important in the merging of and consolidation of KGs and information, thereby enhancing the overall richness of data and allowing for error checking between disparate datasets. Databases such as DBpedia and YAGO3 are rich sources of information to draw from, although the sampling and EA methods used are not always accurate.

7 Background

Many KGs such as DBpedia and YAGO3 have differing design philosophies which can cause issues during the sampling and EA tasks. Both DBpedia and YAGO3 seek to create a large, single multilingual knowledge base (KB) fused from the various Wikipedia KBs of different languages (Auer et al., 2023; Mahdisoltani, Biega, & Suchanek, 2015). However, YAGO3 is an extension of YAGO2, which seeks to create a ”comprehensive anchoring of current ontologies along both the spatial and the temporal dimension” (Hoffart, Suchanek, Berberich, & Weikum, 2013). As such, incorrect sampling and applied EA methods can result in inconsistencies as seen in figure 11 below.

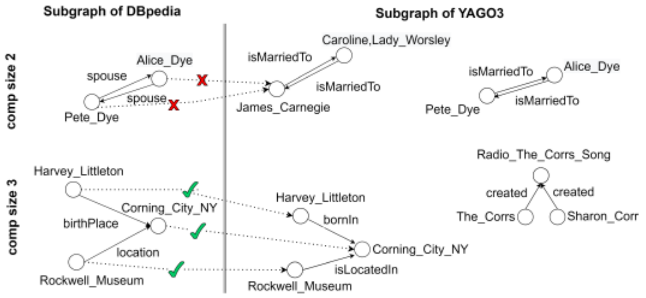
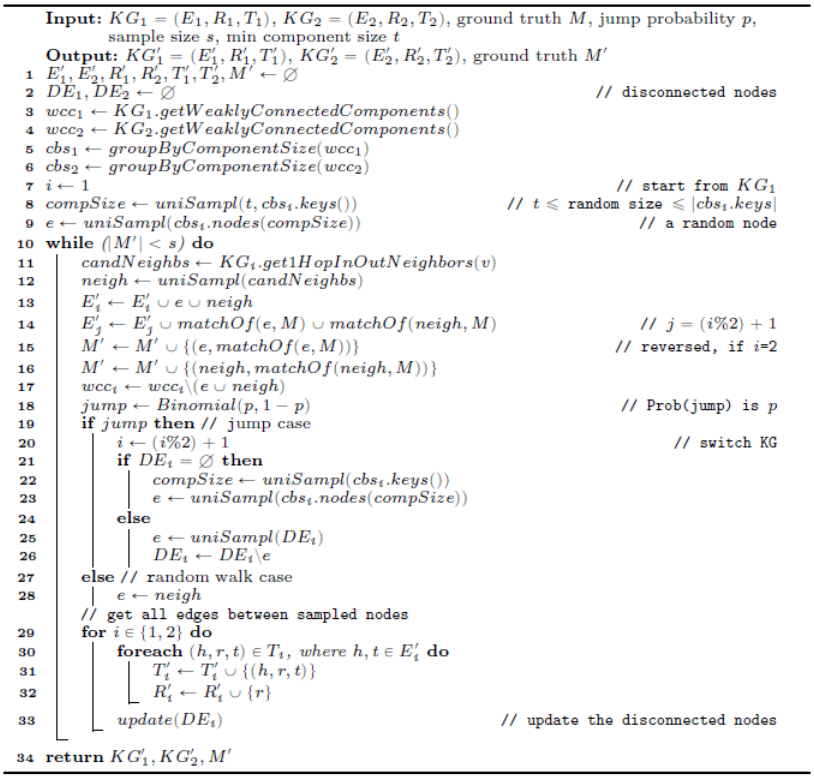


Figure 11: Correct (check) and incorrect (X) matches suggested by RREA (Mao et al., 2020) for nodes belonging to connected components of different sizes on the D-Y dataset.

.

7.1 Exploration-based Sampling Algorithm

Originally a model designed for genetic studies, the Sum of Single Effects model (Wang, Sarkar, Carbonetto, & Stephens, 2020) is adapted as an exploration-based sampling method which utilises a combination of uniform and random walk sampling. It is the first exploration-based sampling method that allows ”controlling (directly) the number and (indirectly) size of connected components of two input KGs” for the EA task. This provides a greater degree of freedom in evaluating the effectiveness of EA methods due to the potential of very large and or very small components affecting the method’s effectiveness over an entire dataset. The sampling algorithm only considers relational information for structural diversity.



```
Input: $KG_1 = (E_1, R_1, T_1)$, $KG_2 = (E_2, R_2, T_2)$, ground truth $M$, jump probability $p$,
       sample size $s$, min component size $t$
Output: $KG'_1 = (E'_1, R'_1, T'_1)$, $KG'_2 = (E'_2, R'_2, T'_2)$, ground truth $M'$
$E'_1, E'_2, R'_1, R'_2, T'_1, T'_2, M' \leftarrow \varnothing$
$DE_1, DE_2 \leftarrow \varnothing$                                  // disconnected nodes
$wcc_1 \leftarrow KG_1.getWeaklyConnectedComponents()$
$wcc_2 \leftarrow KG_2.getWeaklyConnectedComponents()$
$cbs_1 \leftarrow groupByComponentSize(wcc_1)$
$cbs_2 \leftarrow groupByComponentSize(wcc_2)$
$i \leftarrow 1$                                                     // start from $KG_1$
$compSize \leftarrow uniSampl(t, cbs_i.keys())$                      // $t \leqslant$ random size $\leqslant |cbs_i.keys|$
$e \leftarrow uniSampl(cbs_i.nodes(compSize))$                       // a random node
while $(|M'| < s)$ do
    $candNeighbs \leftarrow KG_i.get1HopInOutNeighbors(v)$
    $neigh \leftarrow uniSampl(candNeighbs)$
    $E'_i \leftarrow E'_i \cup e \cup neigh$
    $E'_j \leftarrow E'_j \cup matchOf(e, M) \cup matchOf(neigh, M)$   // $j = (i\%2) + 1$
    $M' \leftarrow M' \cup \{(e, matchOf(e, M))\}$                    // reversed, if $i$=2
    $M' \leftarrow M' \cup \{(neigh, matchOf(neigh, M))\}$
    $wcc_i \leftarrow wcc_i \backslash (e \cup neigh)$
    $jump \leftarrow Binomial(p, 1 - p)$                               // Prob(jump) is $p$
    if $jump$ then // jump case
        $i \leftarrow (i\%2) + 1$                                      // switch KG
        if $DE_i = \varnothing$ then
            $compSize \leftarrow uniSampl(cbs_i.keys())$
            $e \leftarrow uniSampl(cbs_i.nodes(compSize))$
        else
            $e \leftarrow uniSampl(DE_i)$
            $DE_i \leftarrow DE_i \backslash e$
    else // random walk case
        $e \leftarrow neigh$
    // get all edges between sampled nodes
    for $i \in \{1, 2\}$ do
        foreach $(h, r, t) \in T_i$, where $h, t \in E'_i$ do
            $T'_i \leftarrow T'_i \cup \{(h, r, t)\}$
            $R'_i \leftarrow R'_i \cup \{r\}$
        $update(DE_i)$                                                // update the disconnected nodes
return $KG'_1, KG'_2, M'$
```

Figure 12: SUSIE Algorithm

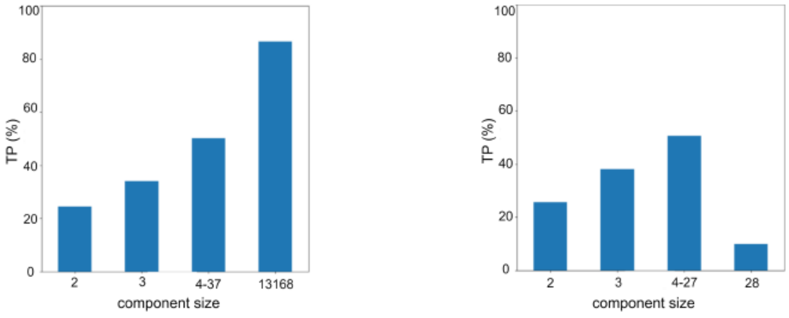
7.2 Process of SuSiE

Initially, the algorithm computes wccs and groups nodes based on their sizes. Then, from one of the input KGs, it performs uniform sampling on components limited by size t as well as a random node belonging to a component of said size. As it samples, a jump probability p is designated which will cause it to jump to the other KG (preferring to jump to a disconnected node) and resume sampling from there. In this way, parts of both input KGs are sampled (explored) almost in concurrence.

8 Method

8.1 Definitions

For measuring the structured diversity of KGs, the paper utilises the following graph-based metrics: ratio of weakly connected components (wccR), max component size (maxCS), average node degree ($\bar{d}$eg), hits@k (H@k), and mean reciprocal rank (MRR). The first three are for measuring jump probability p's impact on sampled datasets, while the latter two are for measuring the EA methods' effectiveness. *wccR* indicates connectivity of a KG: the higher the number of cc, the more sparse the graph is. The inter-connectivity of components and their sizes affects the effectiveness of EA methods, as the biggest connected component largely determines effectiveness of the method for the entire dataset as seen in figure 13.



(a) input ($maxCS = 0.88$)

(b) sample p=1 ($maxCS = 0.03$)

Figure 13: Percentage of true matching pairs (TP) found by RREA for different sizes of connected components on KG_1 of the D-Y dataset. D-Y (a) consists of 15k nodes; its sample (b) consists of 1k nodes.

maxCS is a measure due to the prior explanation. This was determined by early findings in the figure above. The effectiveness of the method is typically lower for smaller connected components. $\bar{d}eg$ is defined as the ratio of total number of incoming and outgoing edges of each node e divided by total number of nodes. Lower average meant high structural diversity. Finally, *H@k* helps determine correctness of the EA method by measuring the proportion of correctly aligned entities ranked in the top k candidates, and *MRR* measures how well it ranks the correct matches.

8.2 Problem Statement

The paper tackles two main topics with regards to the construction and structural diversity of KGs: sampling and EA methods. The paper focuses solely on the implementation of the SuSiE model and how its sampling method affects the utilized EA algorithms.

8.3 Approach

For the datasets, the paper extrapolates four (4) different datasets from six (6) different KGs each with sample size s of 1,000 nodes and minimum component size t set to 1. They test each set on five (5) different p values each being 0, 0.15, 0.5, 0.85, and 1. The four datasets are DBpedia-YAGO3 (D-Y), DBpedia-Wikidata (D-W), BBCmusic-DBpedia (BBC-D), and Memory Alpha-Star Trek Expanded Universe (MEM-E). Each of the datasets focuses on specific entity types and relations (e.g. MEM-E describes TV series related to Star Trek).

For the EA methods, it utilizes embedding-based relational methods RREA, RDGCN, and MultiKE as well as probabilistic, holistic approach PARIS. The code utilised for these methods were from RREA's source[1], OpenEA[2], and entity-matchers[3]; it is unknown whether PARIS source code[4] was used as it is not explicitly stated. As there are two versions of MultiKE and PARIS, the paper utilized only the relational-only version as previously stated due to the fact that the SuSiE model only considers relational information.

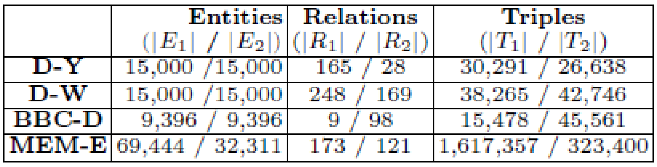
[1] https://github.com/MaoXinn/RREA

[2] https://github.com/nju-websoft/OpenEA

[3] https://github.com/epfl-dlab/entity-matchers

[4] https://github.com/dig-team/PARIS

Figure 14: Datasets Characteristics.



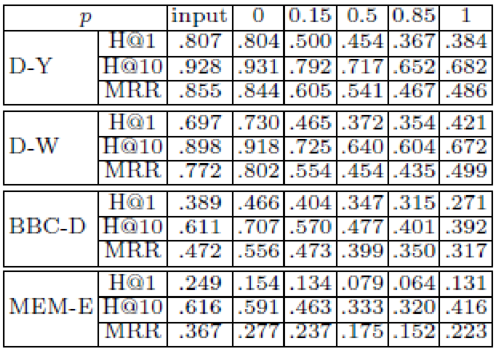
	Entities ($\|E_1\|$ / $\|E_2\|$)	**Relations** ($\|R_1\|$ / $\|R_2\|$)	**Triples** ($\|T_1\|$ / $\|T_2\|$)
D-Y	15,000 /15,000	165 / 28	30,291 / 26,638
D-W	15,000 /15,000	248 / 169	38,265 / 42,746
BBC-D	9,396 / 9,396	9 / 98	15,478 / 45,561
MEM-E	69,444 / 32,311	173 / 121	1,617,357 / 323,400

8.4 Expectations and Observations

9 Results

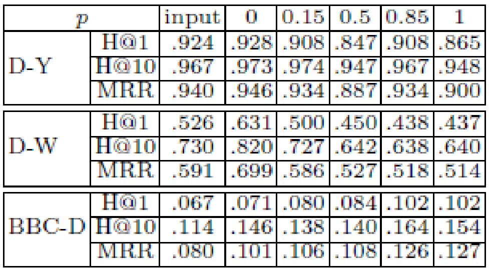
As observed in the tables below, the effectiveness of RREA and MultiKE drop while p increases. This is especially notable when p increases from 0.15 to 0.5 for D-Y and D-W. But for BBC-D, the change in effectiveness does not quite change as drastically for MultiKE despite being overall worse than RREA. Both RDGCN and MultiKE were unable to run on MEM-E (the former due to one-to-one assumption), and the F1-score that PARIS reports is treated as H@1. This was expected as the frequency of jumps led to higher sparsity and more weakly connected components. The effectiveness of RDGCN is largely unaffected by the changes in jump probability, potentially due to set initialization of embeddings with entity names as opposed to RREA's random initialization. Finally, PARIS has a significant change in its F1-score when p increases from 0 to 0.15 due to the fact that it uses only functional relations as opposed to the others which use all relations for embeddings.

Figure 15: Impact of SuSiE on the effectiveness of RREA.



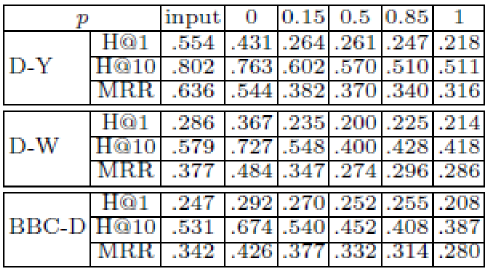
p		input	0	0.15	0.5	0.85	1
D-Y	H@1	.807	.804	.500	.454	.367	.384
	H@10	.928	.931	.792	.717	.652	.682
	MRR	.855	.844	.605	.541	.467	.486
D-W	H@1	.697	.730	.465	.372	.354	.421
	H@10	.898	.918	.725	.640	.604	.672
	MRR	.772	.802	.554	.454	.435	.499
BBC-D	H@1	.389	.466	.404	.347	.315	.271
	H@10	.611	.707	.570	.477	.401	.392
	MRR	.472	.556	.473	.399	.350	.317
MEM-E	H@1	.249	.154	.134	.079	.064	.131
	H@10	.616	.591	.463	.333	.320	.416
	MRR	.367	.277	.237	.175	.152	.223

Figure 16: Impact of SuSiE on the effectiveness of RDGCN.



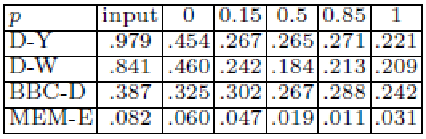
p		input	0	0.15	0.5	0.85	1
D-Y	H@1	.924	.928	.908	.847	.908	.865
	H@10	.967	.973	.974	.947	.967	.948
	MRR	.940	.946	.934	.887	.934	.900
D-W	H@1	.526	.631	.500	.450	.438	.437
	H@10	.730	.820	.727	.642	.638	.640
	MRR	.591	.699	.586	.527	.518	.514
BBC-D	H@1	.067	.071	.080	.084	.102	.102
	H@10	.114	.146	.138	.140	.164	.154
	MRR	.080	.101	.106	.108	.126	.127

Figure 17: Impact of SuSiE on the effectiveness of MultiKE.



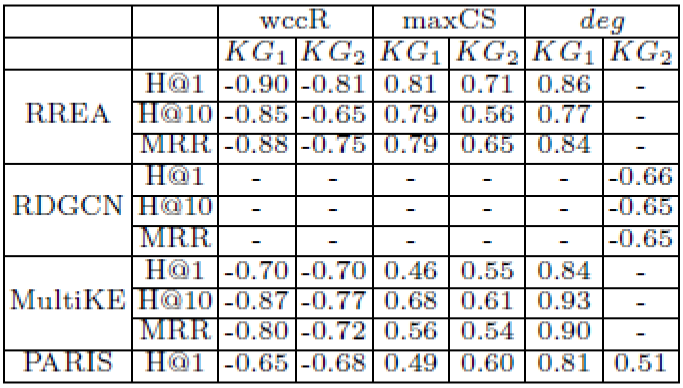
p		input	0	0.15	0.5	0.85	1
D-Y	H@1	.554	.431	.264	.261	.247	.218
	H@10	.802	.763	.602	.570	.510	.511
	MRR	.636	.544	.382	.370	.340	.316
D-W	H@1	.286	.367	.235	.200	.225	.214
	H@10	.579	.727	.548	.400	.428	.418
	MRR	.377	.484	.347	.274	.296	.286
BBC-D	H@1	.247	.292	.270	.252	.255	.208
	H@10	.531	.674	.540	.452	.408	.387
	MRR	.342	.426	.377	.332	.314	.280

Figure 18: Impact of SuSiE on the effectiveness of PARIS.



p	input	0	0.15	0.5	0.85	1
D-Y	.979	.454	.267	.265	.271	.221
D-W	.841	.460	.242	.184	.213	.209
BBC-D	.387	.325	.302	.267	.288	.242
MEM-E	.082	.060	.047	.019	.011	.031

Figure 19: Spearman's correlations between the connectivity of sampled datasets and effectiveness results *Hits@k (H@k)* and *MRR*.



		wccR		maxCS		*deg*	
		KG_1	KG_2	KG_1	KG_2	KG_1	KG_2
RREA	H@1	-0.90	-0.81	0.81	0.71	0.86	-
	H@10	-0.85	-0.65	0.79	0.56	0.77	-
	MRR	-0.88	-0.75	0.79	0.65	0.84	-
RDGCN	H@1	-	-	-	-	-	-0.66
	H@10	-	-	-	-	-	-0.65
	MRR	-	-	-	-	-	-0.65
MultiKE	H@1	-0.70	-0.70	0.46	0.55	0.84	-
	H@10	-0.87	-0.77	0.68	0.61	0.93	-
	MRR	-0.80	-0.72	0.56	0.54	0.90	-
PARIS	H@1	-0.65	-0.68	0.49	0.60	0.81	0.51

10 Conclusion and Future Work

10.1 Conclusions

Although the proposed SuSiE sampling algorithm is not quite perfect in improving the EA methods' effectiveness and reducing their bias, they have shown that structural diversity of EA benchmark data does affect performance of state-of-the-art EA methods. They seek to extend and expand the sampling method to include factual information as well as examine additional diversity measures.

10.2 Related and Future Work

According to the paper, no previous works on KG embeddings had addressed indirect forms of bias related to structural diversity. Such works related to said topic (like Fairwalk) only investigated the impact of direct forms of bias such as in node classification, link prediction, or recommendation tasks. These other graph sampling algorithms were considered, but ultimately the authors settled upon creating and utilising SuSiE as it allows them to ”control the probability with which the output sample will include entities of diverse connected component sizes”.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2023). Dbpedia: A nucleus for a web of open data. In *The semantic web: 6th international semantic web conference, 2nd asian semantic web conference, iswc 2007 + aswc 2007, busan, korea, november 11-15, 2007. proceedings* (p. 722–735). Berlin, Heidelberg: Springer-Verlag. Retrieved from `https://doi.org/10.1007/978-3-540-76298-0_52` doi: 10.1007/978-3-540-76298-0_52

Chiu, J. P. C., & Nichols, E. (2016). *Named entity recognition with bidirectional lstm-cnns.* Retrieved from `https://arxiv.org/abs/1511.08308`

Dsouza, A., Tempelmeier, N., & Demidova, E. (2021). Towards neural schema alignment for openstreetmap and knowledge graphs. In A. Hotho et al. (Eds.), *The semantic web – iswc 2021* (pp. 56–73). Cham: Springer International Publishing.

Dsouza, A., Yu, R., Windoffer, M., & Demidova, E. (2023). Iterative geographic entity alignment with cross-attention. In T. R. Payne et al. (Eds.), *Iswc* (Vol. 14265, p. 216-233). Springer. Retrieved from `http://dblp.uni-trier.de/db/conf/semweb/iswc2023-1.html#DsouzaYWD23`

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013, January). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, *194*, 28–61. Retrieved from `https://doi.org/10.1016/j.artint.2012.06.001` doi: 10.1016/j.artint.2012.06.001

Lin, H., Cheng, X., Wu, X., Yang, F., Shen, D., Wang, Z., ... Yuan, W. (2021). *Cat: Cross attention in vision transformer.* Retrieved from `https://arxiv.org/abs/2106.05786`

Mahdisoltani, F., Biega, J. A., & Suchanek, F. M. (2015). Yago3: A knowledge base from multilingual wikipedias. In *Conference on innovative data systems research.* Retrieved from `https://api.semanticscholar.org/CorpusID:6611164`

Mao, X., Wang, W., Xu, H., Wu, Y., & Lan, M. (2020). Relational reflection entity alignment. In *Proceedings of the 29th acm international conference on information & knowledge management* (p. 1095–1104). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3340531.3412001` doi: 10.1145/3340531.3412001

Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, *3*(4), 333-354. Retrieved from http://jens-lehmann.org/files/2012/linkedgeodata2.pdf

Tempelmeier, N., & Demidova, E. (2021, March). Linking openstreetmap with knowledge graphs — link discovery for schema-agnostic volunteered geographic information. *Future Generation Computer Systems*, *116*, 349–364. Retrieved from http://dx.doi.org/10.1016/j.future.2020.11.003 doi: 10.1016/j.future.2020.11.003

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020, 07). A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(5), 1273-1300. Retrieved from https://doi.org/10.1111/rssb.12388 doi: 10.1111/rssb.12388