Noname manuscript No.
(will be inserted by the editor)

Interpretable Entity Meta-Alignment in Knowledge Graphs using Penalized Regression: A Case Study in the Biomedical Domain

Jorge Martinez-Gil · Riad Mokadem · Franck Morvan · Josef Küng · Abdelkader Hameurlain

Received: date / Accepted: date

Abstract In recent times, the use of knowledge graphs has been massively adopted so that many of these graphs can even be found publicly on the Web. This makes that solutions for solving interoperability problems among them might be in high demand. The reason is that unifying these knowledge graphs could impact a wide range of industrial and academic disciplines that can benefit from aspects such as the ability to configure queries that were not possible until now. For example, in the biomedical domain where there are significant problems of semantic interoperability. To date, several effective methods have been put forward to solve the heterogeneity problems in this knowledge ecosystem. However, it is not possible to assess their superiority in each different scenario they are facing. Therefore, we explore several penalized regression techniques that can mitigate the risk of incurring severe errors in real settings and preserve properties related to the interpretability of the solution. As a result, we have obtained a proposal for entity meta-alignment that yields promising results in the biomedical domain.

Keywords Knowledge Graphs · Knowledge Engineering · Knowledge-based Technology

1 Introduction

A Knowledge Graph (KG) is usually designed to represent any facts in the form of entities and specify how all these entities are connected. It is a powerful approach to connect and unify structured information in a meaningful way and facilitates the design and development of more intelligent information systems. The reason is that it allows both people and machines to consume semantically structured information. This fact has brought meaningful possible solutions for many tasks in

Jorge Martinez-Gil
Software Competence Center Hagenberg, Austria
E-mail: jorge.martinez-gil@scch.at

Riad Mokadem, Franck Morvan, Abdelkader Hameurlain
IRIT Toulouse, France
E-mail: {riad.mokadem, franck.morvan, abdelkader.hameurlain}@irit.fr

Josef Küng
Johannes Kepler University Linz, Austria
E-mail: josef.kueng@jku.at

several computer-related fields, including but not limited to question answering, recommendation systems, information retrieval, and knowledge base completion.

However, when exploiting KGs, some pending challenges remain partially unsolved and therefore require further research. Among all these pending challenges, there is one that stands out from the rest because of the many applications involved: the so-called KG alignment. This challenge aims to find semantic correspondences among multiple KGs. The current trends in this context are focused on entity alignment (EA) since the entities are the potential pivot points that might allow connecting different KGs [38].

Most existing methods typically rely on external information of entities (e.g., Wikipedia entries [26]) and usually require costly manual supervision to complete an alignment. As a result, solutions are needed that improve the process, lower the cost, or at least guarantee some performance when implemented in real systems. For this reason, the EA problem in the context of KGs has received much attention lately due to its great usefulness in a significant number of applications related to the discovery of semantic correspondences between different KGs. For example, it is immediately evident that EA methods are essential to finding the connection points that allow the mapping and merging of graphs that could greatly benefit the existing techniques for query expansion.

Nowadays, there are many different approaches to face the EA challenge. The main problem is that most of the existing techniques are based on the so-called KG embeddings techniques [9]. These techniques are mathematical methodologies to project the entities in a continuous vector space, so these embeddings can be suitable to be efficiently processed by deep neural networks. This fact is both a great advantage and a significant disadvantage of such approaches. On the one hand, neural solutions can achieve the best results in most EA tasks [29]. However, as they are based on neuronal learning, these models act as a black box, i.e., there is no human operator who can understand how they work and need vast amounts of data to be trained with a certain guarantee of success [7].

Our proposal focuses on a radical alternative to the use of embeddings. We want to take advantage of the fact that the research community has proposed many different interpretable strategies of a very different nature to face this challenge. Most empirical studies show us is that it is not possible to determine which of these strategies is the best since this often depends on the context, the type of graph, and even the use case on which the alignment challenge is focused. Our hypothesis is that it should be possible to design a meta-alignment[1] of interpretable methods that can be put into operation in real settings with certain guarantees in terms of performance. Therefore, the contribution of this work is:

- We propose a new approach for entity meta-alignment so that an intelligent aggregation of existing methods can be put into production. The idea behind this proposal is to explore some penalized regression methods to facilitate better generalization and, therefore, benefit from the great degree of precision of some existing methods and, at the same time, the great coverage that can be expected from other methods. In addition, we put as a restriction that we cannot use black-boxes strategies as inputs of the ensemble.
- Moreover, under the assumption that a high degree of interpretability is better than a few hundredths of precision, we evaluate and compare our proposal with existing solutions using several popular benchmark datasets to show that putting an ensemble of methods into production can provide more significant guarantees than relying on just one single method in terms of accuracy and interpretability in real KG alignment scenarios.

[1] The term meta-alignment reflects the idea that we do not build an alignment method from scratch, but we aim to build an ensemble of existing ones

The rest of this paper is structured as follows: In Section 2, we present the state-of-the-art concerning entity alignment in the context of KGs. In Section 3, we introduce our novel strategy for entity meta-alignment. In section 4, we present the results we have obtained after evaluating our proposal regarding different biomedical datasets and an appropriate comparison with methods representing the state-of-the-art. Finally, we conclude with the lessons learned as well as the future lines of research.

2 State-of-the-art

The use of cognitive resources for their exploitation by automatic systems has gained much attention lately. The idea behind this is that they should serve as a data-driven framework in research related to cognitive activities, and therefore it acquires direct importance in Artificial Intelligence. Nowadays, these postulates have been possible thanks to the creation and publication of enormous amounts of data stored and linked in the KGs. This knowledge is often accessible through the Web, which provides a common framework for sharing and reusing data across application, enterprise, and community boundaries.

A KG is a knowledge base in the form of a graph that has been designed specifically to meet the complex storage and retrieval requirements of automated knowledge management with the support of Artificial Intelligence or expert systems. Due to the great importance of social sciences and communication networks such as social media, KGs have expanded in scale and popularity. With the advent of these huge datasets, there is a growing need to provide big-scale analysis methods and tools that can facilitate their exploitation. This is why methods to deal with large KGs have attracted much attention in recent times. KGs usually consist of a set of entities and information about the relationships between those entities. This knowledge can then be used to improve a large number of existing methods in various computer-related problems. For example:

- Question Answering systems, which are systems that try to reply in the most natural way possible to questions asked by people. For example, the popular IBM Watson is a big-scale question answering system that uses several knowledge bases in the form of a graph such as YAGO and DBpedia as data sources to answer a wide range of questions of general-purpose [6].
- Recommender Systems, which are computer systems that attempt to provide specific content recommendations for the users. The fact is that the vast volume and variety of online content such as books, movies, and news has become a severe problem for users. The proper exploitation of KGs can help to improve the recommender accuracy and increase the diversity of recommended items. For instance, DKN [33] is a method based on CNN proposed to use KGs for news recommendation.
- Information Retrieval, whereby we can see the example of Wikidata, which is a free and collaborative KG that collects structured data to support Wikipedia, the Wikimedia Commons, and the other wikis from the Wikimedia ecosystem [32]. Its great advantage is the imposition of a high degree of structured organization, which allows easy retrieval of data by the Wikimedia projects and determines how it can be reused.
- Knowledge Base Completion, which is one of the main forms of reasoning that is sought to be carried out on knowledge bases. Since the facts in a knowledge base are stored in the form of triplets of the form (subject, predicate, object), the completion objective is to guess new triplets [21].

In addition, another of the significant advantages of KGs is that they allow the integration of new knowledge from multiple external KGs. Unfortunately, KGs are usually built using different languages or with diverse ontological foundations, which results in great difficulty in integrating the knowledge of external KGs [1]. For this reason, the EA challenge arises in the context of the KGs. The task of EA alignment goes back to the challenge of ontology matching, whereby it was a matter of finding the semantic correspondences between ontologies of the same domain in an unsupervised way [4]. While the problem of EA in KGs has evolved by researching as broad vocabularies as possible and establishing them as a standard, the most recent approaches take a more data-driven view. Recently, the research community has become actively involved in the design and development of embedding techniques. Nowadays, most KG alignment solutions rely on embeddings to determine similar elements in different KGs [31].

Some of the most relevant works in this area include semantic aggregation and attribute attention [11], Trisedya et al. propose to use attribute embeddings [30]. In contrast, Yang et al. propose a co-training schema that considers both structures and attribute embeddings [36]. Berrendorf et al. suggest using Graph Convolutional Networks [2] or Wu et al. that propose a relation-aware strategy for EA in heterogeneous KGs [34], and many more.

The major problem here is that mapping information from different KGs is far from being trivial. In a typical setting, some of the alignments are known in advance (seed alignments), and the task is therefore supervised. However, most of the time, this seed is not available, and therefore it is necessary to rely on transfer learning techniques to overcome the limitations of a cold start. In recent times, some research has proven that to obtain high-quality alignments. It is crucial to combine information from different sources [29].

Furthermore, such a combination must be done strategically rather than following a short-sighted approach. To do that, it is necessary to make some mild assumptions around the idea of those equivalent entities in multiple KGs might usually have similar neighbors and attribute names. Several researchers have already formulated this hypothesis. It is naturally assumed that a system for EA could be either:

- Based on the KG structure. These methods are good for aligning due to their simplicity, generality, and ability to deal with large-scale data. Initially, they utilize KG representation methods to represent structural properties and embed KGs into individual low-dimensional spaces. In this way, it is possible to extract complex features that take advantage of the knowledge graph topology or leverage multi-step connections between entities hidden by simple text analysis.
- Based on the entity attributes. For example, alignment of entities based on similarity. Older entity alignment approaches use string similarity as the primary alignment tool. For instance, LIMES [22] uses the triangle inequality to calculate correspondences. Current methods mainly capture semantic information of entities, as semantics can be easily encoded as embeddings, facilitating the fusion of different feature representations.

It is usual to consider some weighted distance function to properly balance both aspects, which combines structural embeddings and entity's attribute embeddings. For example, Eq. 1 shows a method for entity alignment prediction; both features could be defined as:

$$Dist(e_{i1}, e_{i2}) = \theta Dist'(e_{i1}, e_{i2}) + (1 - \theta) Dist''(e_{i1}, e_{i2}) \tag{1}$$

Where θ is a hyper-parameter balancing the importance of both kinds of embeddings, however, solutions of this kind have two fundamental disadvantages: they require vast amounts of data to be trained, and they are not interpretable by a human operator.

However, in practice, it is also possible to reach that balance using alternative methods. Before embeddings became popular in the community, many alignment methods worked decently well and were easy to understand by a human operator. This is especially interesting in the biomedical field where one tends to work with knowledge models that are really large [18]. These alignment techniques used to be divided into two: Knowledge-based approaches and corpus-based approaches.

Knowledge-based approaches differ from the corpus-based approaches relying on co-occurrence or distributional similarity. Knowledge-based approaches are usually considered when working with structural KGs [37], while corpus-based approaches are typically applied in textual corpora [28]. Several works mainly exploit the concept level knowledge (similar to the conceptual schema in database setting), while the instance level knowledge is used to support the concept of knowledge. Our research can focus on both kinds of entities. Our idea is to use these classical techniques to add them strategically into an ensemble of methods that can do EA tasks with great precision but without giving up interpretability. Besides, the volume required to train our technique is several orders of magnitude less than that of neural network-based methods. We will explain the technical details of our approach below.

3 Methods for Entity Meta-Alignment between Knowledge Graphs

Despite the large number of alignment techniques that have been proposed to date, finding the most suitable alignment approach is still an issue since it is challenging to discern which of them should be used in each situation. For this reason, we aim to find an appropriate technique to reduce the tedious task of creating entity alignments manually in the best possible way. So we propose a method for entity meta-alignment that reduces the uncertainty when implementing an exploitation strategy. As we have already mentioned, there are currently many techniques for addressing the challenge, so we aim to aggregate them following a strategic way to reduce the risk of making mistakes in real scenarios.

One thing to consider is that concepts of the KG contain axioms describing concept hierarchies and are usually known as classes (TBox). In contrast, axioms about entity instances are usually known as instances (ABox). Therefore, one part is dedicated to the terminological definition of concepts and their relationships, and the other is dedicated to instances (some authors also use the term resource). While working on the TBox is a problem that can be solved from a wide range of solutions since concepts are connected through relationships, and it is possible to calculate paths, give weights to the connections, etc. working with the Abox is much more problematic because they usually do not have that advantage. In Fig. 1, we can see an example of a) the segmentation between TBox and ABox, and b) the opportunities that a good alignment offers. Being able to find good pivot points invariably leads to the possibility of automatically acquiring new knowledge. In fact, by discovering the optical organ - eye many new facts can be inferred and the resulting model is considered to be more complete.

To find the correspondences, we aim to, instead of combining the features that come from the structure with those that come from the entities; our proposal is based on a higher level of aggregation. The units to be added are already existing methods. In addition, such an addition can be made in many different ways. However, we consider traditional operators such as arithmetic, geometric and harmonic media, or median, or mode are trivial. We intend to make such aggregation using different kinds of regression as more complex interactions can be modeled and remain interpretable.

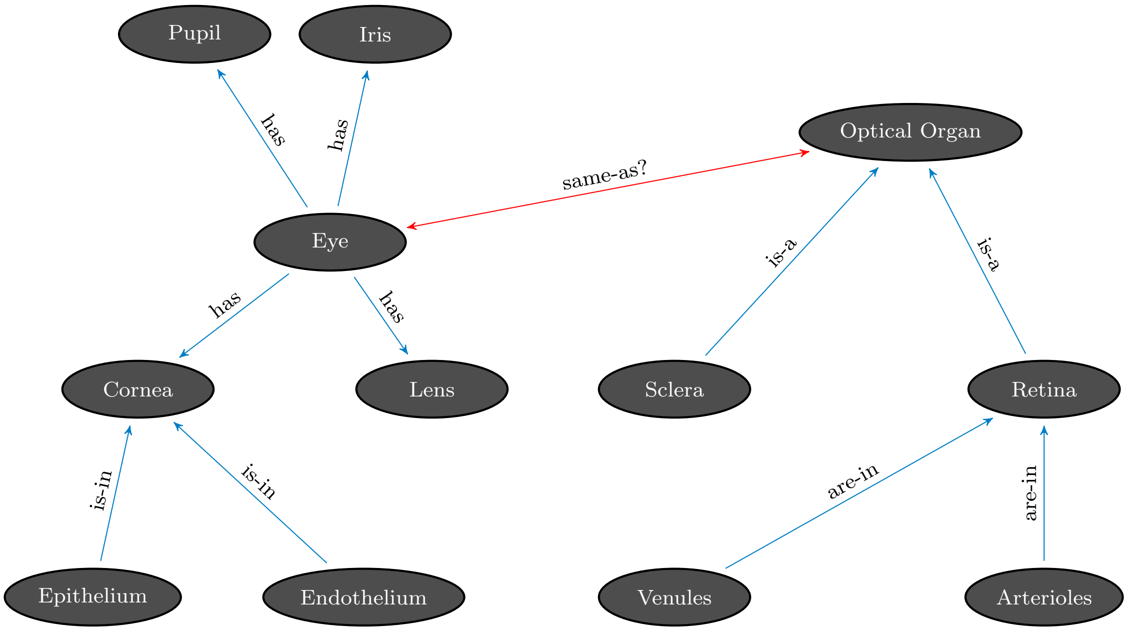

Fig. 1 Example of alignment between KGs. The possibility of finding a pivot point between the entities optical organ and eye can facilitate the obtaining of new knowledge that can be useful in a wide range of domains

On the other hand, in this work, we rely on one of the most popular general-purpose knowledge bases, i.e., WordNet[2] which is a knowledge base that attempts to model synonymic relationships as well as sub summary relationships between concepts. One of the most important characteristics of this modeling method is that knowledge is structured in the form of a hierarchy. There are several methods for calculating similarity based on the different paths calculated in such a taxonomy. These methods are Path [24], Leacock [16], Wu & Palmer [35], Li [19], Resnik [25], Lin [20], Jiang & Conrad [12], and wpath [37]). Our concept here is to aggregate all these methods strategically so that a) we can reach higher levels of accuracy b) any person could take a look at WordNet and the way the distance between the different concepts is calculated to realize where the final semantic similarity value comes from. Below, we explain how we propose to build the different ensembles.

3.1 Linear Regression

First of all, it seems like a good idea to explore linear models. Linear regression is one of the most basic yet most interpretable approaches to aggregate numerical inputs. In fact, in linear regression scenarios, it is possible to provide several inputs and, in return, get a meaningful value as output. This is done by giving some weights to each of the input variables. In its most basic form, linear regression does not penalize for its choice of weights. More formally, we look for a function:

$$y = \alpha + \beta x \tag{2}$$

so that we can find

[2] https://wordnet.princeton.edu/

$$\min_{\alpha,\beta} Q(\alpha,\beta), \quad \text{for} \quad Q(\alpha,\beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \ . \tag{3}$$

In this way, we want to minimize the error, i.e., the absolute measure of the shortest distance that the points fall from the regression line. However, our hypothesis is that such a simplistic model might not work very well. Since the number of simple methods to be added is too large and the solution may not detect the most suitable methods. For that reason, we considered using some types of penalized regression. These types of regression allow us to create models penalized for having too many variables in the model, facilitating regularization and making over-fitting difficult.

3.2 Lars Regression

Although linear regression is well known, it is often considered too simple to work well in real environments. Thus, some variants have been designed that usually work better since they have more sophisticated error regularization methods. For this reason, we could consider least-angle regression (Lars) as an algorithm for fitting linear regression models to input data. The consequence of imposing a penalty is to reduce the coefficient values towards zero. This allows the less contributive variables to have a coefficient close to zero. This kind of regression is used when over-fitting is a concern (in our case, since we are concerned about our ensemble's prediction capability). More formally,

$$y = \alpha + \beta x, \ \min_{\alpha,\beta} \left\{ \sum_{i=1}^{N} (y_i - \alpha - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \tag{4}$$

The idea here is to identify the variable most correlated with the response instead of fitting that variable, trying to move its associated coefficient iteratively towards its least-squares value. Although the concept is simple, it usually works quite well in practice.

3.3 Neural regressors

However, the results that can be obtained with the above technique are not usually optimal. For this reason, an alternative way to address the problem is to build a regressor for meta-alignment based on neural networks. This regressor must be able to take the inputs from existing alignment algorithms and must be able to configure a neural network whose output is the value of the meta-alignment. This way of working could be able to guarantee that no single algorithm is relied upon in production environments. We are interested in Multi-Layer Perceptron (MLP), a class of feed-forward artificial neural networks. More formally, we need to find the weights, so that:

$$y_i = f(x_i, w) = w^T x_i \tag{5}$$

In this way, we can calculate the distance to the error E.

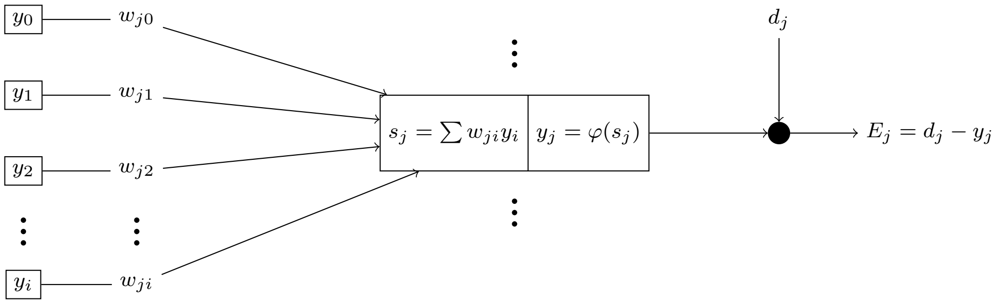

Fig. 2 Example of Multi-Layer Perceptron whereby a number of inputs are aggregated using a neuron, and the desired output is compared to the aggregated scored in order to obtain the error

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(f(\mathbf{x}_i, \mathbf{w}) - y_i\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 \tag{6}$$

In Fig. 2, we can see an example of how this kind of regressors try to minimize the error between what is perceived as the desired result d_j and the result obtained through the network y_j.

The great advantage of MLP is its capability to learn non-linear models. The problem is that although these types of regressors usually work pretty well, they are challenging to interpret. That means that we could give them the inputs and get a result, but a human operator would not understand what happened in the middle of the process. When working with only one neuronal layer, such as this case, the interpretation is still possible. In addition, they require large amounts of data to be trained appropriately. However, there are no significant amounts of labeled data in this context, so their application is often not optimal. Last but not least, the process of initially working random weight leads to different results in each execution..

3.4 ElasticNet

Although working with neural networks is more likely to yield better results, there is a small problem in their interpretability. It is possible to train the neural network, give it the inputs and get an output that is usually quite accurate. However, a person cannot understand what happens within that neural network since the only observable thing is many nodes and connections between them. For this reason, sometimes it can make sense to use other strategies that, although not so precise, are easier to interpret by a human operator. Therefore, as a possible ensemble, we also consider ElasticNet, which aims to fit a linear model with coefficients $w = w_1, w_2, ..., w_n$ to minimize the residual sum of squares between the observed targets in the dataset. This regressor is appropriate in the fitting of linear regression models, and therefore it is highly interpretable. The ElasticNet is a regularized regression method that linearly combines the L1 and L2 penalties of other methods. It is a regressor with several advantages that include but is not limited to reducing the number of predictors in the regression model by identifying the most important predictors and discarding the redundant predictors. The great advantage of ElasticNet is that it can estimate with potentially fewer predictive errors than the classical least squares method. So it is ideal when dealing with highly correlated variables, so that is our case (The algorithms we use as input are already quite accurate) since highly correlated inputs will tend to have similar coefficients. More formally, ElasticNet regression can be defined as:

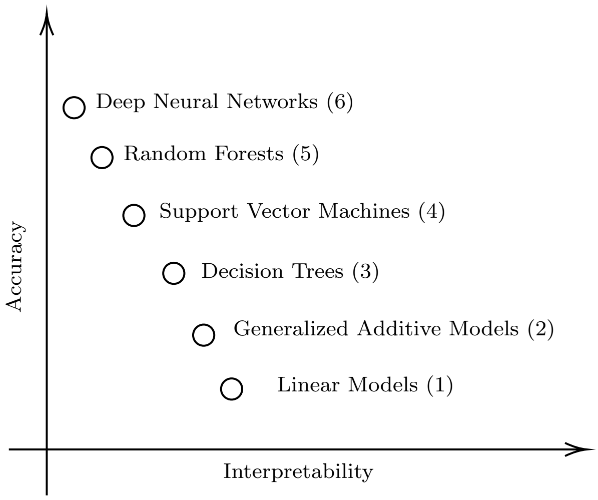

Fig. 3 Classification of the different ML methods with regards to their trade-off between accuracy and interpretability

$$\hat{\beta} = \text{argmin}(\|y - X\beta\|^2 + \alpha_2\|\beta\|^2 + \alpha_1\|\beta\|_1). \tag{7}$$

ElasticNet requires us to tune parameters to identify the best α and β. The concept behind this regressor is that it penalizes the sum of absolute values of the weights and the sum of squared value of the weights being regulated with another coefficient. As a result, highly correlated inputs are assumed to have similar estimated coefficients. Our hypothesis is that this method should work well in our scenario since aggregating methods that are already quite good, i.e., individual results, are somehow similar.

3.5 Considerations on the Interpretability

Interpretability has gained much specific weight in recent years. This is since during the last decade, there has been a constant race to build models that yield better and better results, but this race has lost sight that in the end, models must be used by people. Models will not be advantageous if the people who use them cannot look at them, study them and understand them. If this is not the case, it is clear that there will be great distrust and aversion.

The truth is that while there are methods that have very high levels of interpretability: Consider a regression model (with or without penalty) on a set of basic semantic similarity measures. However, empirical evidence shows that easily interpretable models are often quite simplistic, so one must move towards methods capable of complex modeling interactions between algorithms in the ensemble. Our approach considers that, even then, these interactions must be easily understandable and reliable.

Figure 3 shows a highly accepted classification of the different methods concerning the trade-off between accuracy and interpretability that they can model. The numerical values indicate the level of interpretability that is usually associated with each of the different strategies. As can be seen, the methods close to regression are highly interpretable since they allow us to precisely know how each of the input variables is involved in producing the output value. On the contrary, the deep neural solutions do not easily find out this correspondence and are therefore considered black boxes in practice. In the frame of this work, we will use atomic methods based mainly on the use of dictionaries. These methods have been used recurrently for many years. They are highly interpretable since any human operator can understand how to calculate the similarity between words based on their location in a computer dictionary.

4 Experimental Studies

In this section, we present the results we have obtained after evaluating the performance of our approach using large KGs. To do that, we have organized the section to explain the biomedical benchmarks datasets to be used, the evaluation criteria that we will follow, the requirements for parameter selection and cross-validation, the empirical results that we have achieved, and the discussion on the results that we have achieved.

4.1 Datasets

In the context of this work, we will try to deal with domain-specific benchmark datasets from the biomedical field, which are challenging to work with in terms of volume and variety. This will give will us an idea about the behavior of our ensembles in real environments. The datasets used by entity alignments methods are generally based on large-scale open-source data sources. In the scope of this work, we have focused on KGs from the biomedical domain already published by Kolyvakis et al. [13].

We consider here the foundational Model of Anatomy (FMA), a KG that represents the phenotypic structure of the human body [23]. The Adult Mouse Anatomical Dictionary (MA), which is a KG representing the anatomy of an adult mouse [10]. The NCI Thesaurus (NCI) provides a standard terminology for cancer [3] and its anatomy subdomain describe naturally occurring human biological structures. And finally, the SNOMED collection (SNOMED), which is a KG of medical terminology to be used in clinical reports [5]. Some interesting examples of similar entities that we need to identify are shown in Table 1. As can be seen in the sample, some correspondences are immediate due to their lexical similarity. However, other correspondences are not trivial and require the use of background knowledge to solve them.

FMA	**NCI**	**Correspondence**
triceps surae	triceps surae	True
endometrial cavity	cavum uteri	True
oropharynx	human pharynx	True
fallopian tube	uterine tubes	True
mammotroph	lactotrope cell	True
skull bone	human cranium	True

Table 1 Some examples of positive correspondences between two biomedical KGs

One of the characteristics in this context is the abundance of positive samples. The reason is that, in real life, the experts usually provide positive correspondences between KGs. There is no point in indicating negative cases, as negative cases are all those that are not positive. However, this abundance of positive samples makes difficult the learning process, as most regression ensembles assume the processing of balanced datasets (datasets with a similar number of positive and negative samples). In practice, the number of similar entities between two KGs is several orders of magnitude smaller than the number of all possible combinations.

	entities1	entities2	#positives	#possible correspondences
MA-NCI	2,744	3,304	1,489	9,066,176
FMA-NCI	3,696	6,488	2,504	23,979,648
FMA-SNOMED	10,157	13,412	7,774	136,225,684

Table 2 Summary of the biomedical graphs we consider for our empirical studies

For this reason, we use the ground truth together with a randomly generated amount of negative samples. This technique is well known in machine learning and consists of the random and partial corruption of correct input data. For that reason, we aim to avoid an imbalance problem in the context of this work.

Moreover, to realize the size of the datasets we are working with, we show in Table 2, where it can be clearly seen that the proportion of corresponding entities is tiny compared to the total number of possible comparisons that can be established. So we are faced with a complicated problem.

In addition, it should be noted that there are many correspondences between those KGs that are almost similar as we have already seen. This is very common when working with real benchmarks, whereby everybody uses the exactly same names. For cases like this, it is very informative to also provide information regarding the Levenshtein distance [17]. This distance is considered to be a baseline in the literature since the idea behind is simple but effective: the similarity between two (sets of) words is the minimum amount of single-character edits. More formally, the similarity between two words could be considered as the inverse of the distance, whereby the distance is defined as:

$$\text{d(i,j)} = \begin{cases} \max(i,j) & \text{if} \min(i,j) = 0, \\ \min \begin{cases} \mathrm{d}(i-1,j)+1 \\ \mathrm{d}(i,j-1)+1 \\ \mathrm{d}(i-1,j-1)+1 \end{cases} & \text{otherwise} \end{cases}$$

The simplicity of this method does not prevent it from being one of the most used in practice. Since most semantic correspondences that are usually found in real scenarios only differ in a few characters.

4.2 Evaluation criteria

Experts do not usually provide values within a real numerical scale for the correspondence between entities in real life. Such information is often confusing as well as tedious and challenging to obtain. Therefore, experts usually provide their knowledge regarding the correspondence of entities using a binary classification: the entities being compared are equivalent or not. Therefore, this type of problem is usually evaluated by measuring the percentage of cases that the machine has found out correctly.

As we have already mentioned, there is a way to evaluate a set of alignments between KGs in which correct and incorrect mappings are identified to use the trained classifier to predict whether an assertion of semantic equivalence between two concepts is or is not valid. To evaluate the accuracy based on the ratio of successful cases, we need to use the precision as the fraction of retrieved mappings that are relevant to the process where obviously, the greater the number of relevant correspondences (i.e., the lower the number of incorrect cases), the precision will take a value closer to 1 (or 100% if we want to work with percentages).

Method	ρ	interp.
path	91.2%	0
lch	94.0%	0
wup	91.8%	0
li	91.5%	0
res	95.8%	0
lin	93.1%	0
jcn	90.9%	0
wpath	91.6%	0
OM-LSTM	97.0%	6
OM-TBERT	97.7%	6
OM-LSTM + SGAT	97.5%	6
OM-TBERT + GraphSAGE	95.4%	6
OM-TBERT + TransE	89.0%	6
DAEOM	98.1%	6
DOME	99.3%	6
LR	99.5%	1
Lars	99.5%	1
ElasticNet	99.5%	1
MLP	99.7%	2

Table 3 Results obtained for the FMA-NCI

4.3 Parameter selection and cross-validation

It is necessary to note that over-fitting is a common problem that can occur in most trained models. To avoid that, k-fold cross-validation can be performed to verify that a given model is not over-fitted. In this work, all our models are cross-validated. In addition, we consistently report the highest cross-validation value. Furthermore, the implementation for the atomic methods is based on Sematch[3]. However, some adaptations had to be made as the Wordnet dictionary is not able to work with such specific terms. Therefore, we have transformed these models into a bag-of-words model with individual word distances.

On the other hand, the implementation for the regression methods is based on Scikit-Learn[4]. It is necessary to point out that the default settings of the mentioned methods have been used in all the reported experiments. Moreover, the comparisons will be performed using the following methods: DAEOM [14], DOME [27], and OM [34]. The reason for our decision is that these methods have been the ones that have achieved the best results in the biomedical field so far.

4.4 Performed experiments

Below, we show the experiments we have conducted. Most of these algorithms cannot work with short textual expressions of the "membrane of smooth endoplasmic reticulum" type. As we have already mentioned, they use general-purpose dictionaries to calculate distances. But it is possible to segment such short textual expressions and calculate the similarity between all words and then calculate an average. In this way, such algorithms would also be prepared to work in this scenario.

Table 3 shows the results for the benchmark FMA-NCI that contains 2772 positive labeled cases. Even though the results may seem very high, it is necessary to remark that 1684 of 2772 cases (60.75%) are trivial and could be solved using the Levenshtein distance.

[3] https://gsi-upm.github.io/sematch/
[4] https://scikit-learn.org/

Method	ρ	interp.
path	87.9%	0
lch	96.7%	0
wup	92.7%	0
li	90.3%	0
res	96.7%	0
lin	92.7%	0
jcn	86.9%	0
wpath	90.0%	0
OM-LSTM	95.8%	6
OM-TBERT	96.6%	6
OM-LSTM + SGAT	97.1%	6
OM-TBERT + GraphSAGE	98.1%	6
OM-TBERT + TransE	96.1%	6
DAEOM	98.9%	6
DOME	98.5%	6
LR	99.7%	1
Lars	99.7%	1
ElasticNet	99.7%	1
MLP	99.8%	2

Table 4 Results obtained for the FMA-SNOMED

Method	ρ	interp.
path	91.1%	0
lch	96.0%	0
wup	92.2%	0
li	91.6%	0
res	96.0%	0
lin	91.2%	0
jcn	90.5%	0
wpath	91.1%	0
OM-LSTM	97.2%	6
OM-TBERT	97.7%	6
OM-LSTM + SGAT	98.1%	6
OM-TBERT + GraphSAGE	91.3%	6
OM-TBERT + TransE	72.2%	6
DAEOM	99.0%	6
DOME	98.8%	6
LR	99.0%	1
Lars	99.0%	1
ElasticNet	99.0%	1
MLP	99.1%	2

Table 5 Results obtained for the MA-NCI

Table 4 shows the results for the FMA-SNOMED benchmark that contains 8503 positive cases. Again, it is not appropriate to look solely at the absolute values for accuracy since 4755 of 8503 cases (55.92%) are trivial and could be solved with a simpler method.

Finally, Table 5 shows the results for the MA-NCI that contains 1496 positive cases. Once again, it is necessary to mention that for the MA-NCI benchmark, 729 of 1496 cases (48.73%) are trivial and could be solved by solely using the Levenshtein distance. Moreover, it is clear that as the samples available for training become much more numerous, the neural regressor begins to demonstrate its superiority regarding accuracy, although not in terms of interpretability.

When the training data sets are enormous, the neural regressor can beat the rest. Because while others assume linear relationships between data, the neural regressor can also detect non-linear relationships, which gives it a clear advantage. However, penalized regression methods obtain outstanding results and are much easier to interpret since the mapping between the inputs and output is explicitly stated. So it is up to the user to decide whether it needs more accuracy at the expense of the interpretability of the resulting model or vice versa.

4.5 Discussion

One of the most critical aspects of the KGs is the quality of the represented information, including other KGs. This is precisely at this point that the importance of having suitable alignment methods comes into play. As we have seen, there may be times when the best idea is not to develop an alignment strategy from scratch but to design a solution that allows for the effective aggregation of existing strategies. Many similar experiences already exist, for example, in the world of word embeddings, where recent research can demonstrate that the linear combination of existing methods is capable of surpassing the state-of-the-art [15]. It is also the case of meta-matching [8], where traditional matching algorithms are added in a strategic way to solve the problems that usually affect simple strategies when it comes to their exploitation.

From the results obtained from our experiments, it is worth noting that the neuronal regressor gets the best results. Nevertheless, it is very convenient to point out two facts: a) it requires large amounts of training data, and b) it is more difficult to interpret than other methods, for example, the ElasticNet regressor, which does allow visualizing how the input measurements are aggregated to shape the output. Moreover, by selecting the most relevant predictors and discarding the redundant ones, the ensemble in the regression model is reduced which helps it to generalize better. We can also observe that unlike synthetic datasets designed to test complex cases in real KG alignment cases, there are many trivial situations to solve, using, for example, Levhenstein's distance. In our domain-specific benchmarks, we have found that almost half of the similarity cases identified by the domain expert were trivial, so the accuracy of the results was usually relatively high. Therefore, the interesting part is to observe how the methods studied behave for not trivial cases.

Last but not least, although it may seem that models of increasing complexity will always perform better, empirical evidence shows that easily interpretable models are not always simplistic, so sometimes it is not necessary to move towards methods capable of complex modeling interactions between methods in the ensemble. In fact, working with an intelligent aggregation of simple methods that are already quite good has proven to be sufficient in the context of this study.

5 Conclusions

In this work, we have presented our novel proposal for the meta-alignment of entities in the context of KGs. To do so, and unlike most proposals in this context, instead of designing and developing yet a new EA method, we propose a higher-level technique capable of performing a strategic aggregation of an already existing pool of methods of diverse nature. To this end, we have studied various forms based on penalized regression. In this way, if an entity alignment method does not behave too well for a specific case, its impact can be mitigated by the rest of the alignment methods. Hence, as a final result, there are certain guarantees that there will be no significant errors when putting the system in operation in a real environment.

As we have seen in our evaluation, the methods based on penalized regression can beat most existing solutions. The reason is that such aggregation strategy can benefit from the precision of current techniques and at the same time dilute their weaknesses. To do that, our approach tries to find the best way to add the semantic similarity values predicted by multiple existing methods to determine the correspondences between the entities of the knowledge graph. We have also seen that, although methods based on neural networks may give better results, they require vast amounts of data for training and are challenging to interpret by a human operator, so regressors such as ElasticNet could make sense in situations whereby a high degree of interpretability is better than a few hundredths of additional accuracy.

As future work, we have to work to discover other types of relationships between entities. In this work, we have developed a strategy to discover same-as-type relationships. However, there are many more types of relationships possible. For example, WordNet models its knowledge with up to 18 different types of relationships, so we should train our models to recognize more types of relations between entities. Secondly, we propose a multi-objective optimization scheme. So that the meta-alignment proposal can be configured to fit the user's need; thus, it is necessary to calculate a Pareto front of solutions that cover two opposing objectives: accuracy and interpretability. So the user can select the type of solution that best fits the scenario to be solved. In this work, interpretability is inherent in the ensembles that we use, but we cannot control it. With a multi-objective strategy, we could give the user the possibility to define what levels of interpretability and accuracy are tolerable.

Acknowledgments

We would like to thank the anonymous reviewers for their help in improving the manuscript. This work has been funded by the project FR06/2020 by International Cooperation & Mobility (ICM) of the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH). It is also supported in part by the French Ministries of Europe and Foreign Affairs (MEFA) and of Higher Education, Research and Innovation (MHERI). Amadeus Program 2020 (French-Austrian Hubert Curien Partnership – PHC). Grant Number 44086TD.

References

1. A. Agibetov and M. Samwald. Benchmarking neural embeddings for link prediction in knowledge graphs under semantic and structural changes. *Journal of Web Semantics*, 64:100590, 2020.
2. M. Berrendorf, E. Faerman, V. Melnychuk, V. Tresp, and T. Seidl. Knowledge graph entity alignment with graph convolutional networks: Lessons learned. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 3–11. Springer, 2020.
3. S. de Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright. NCI thesaurus: Using science-based terminology to integrate cancer research results. In M. Fieschi, E. W. Coiera, and J. Y. Li, editors, *MEDINFO 2004 - Proceedings of the 11th World Congress on Medical Informatics, San Francisco, California, USA, September 7-11, 2004*, volume 107 of *Studies in Health Technology and Informatics*, pages 33–37. IOS Press, 2004.
4. H. H. Do and E. Rahm. COMA - A system for flexible combination of schema matching approaches. In *Proceedings of 28th International Conference on Very Large Data Bases, VLDB 2002, Hong Kong, August 20-23, 2002*, pages 610–621. Morgan Kaufmann, 2002.
5. K. Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.
6. D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.
7. J. M. Gil and J. M. Chaves-González. Automatic design of semantic similarity controllers based on fuzzy logics. *Expert Syst. Appl.*, 131:45–59, 2019.
8. J. M. Gil and J. F. A. Montes. An overview of current ontology meta-matching solutions. *Knowledge Eng. Review*, 27(4):393–412, 2012.
9. Y. Hao, Y. Zhang, S. He, K. Liu, and J. Zhao. A joint embedding method for entity alignment of knowledge bases. In H. Chen, H. Ji, L. Sun, H. Wang, T. Qian, and T. Ruan, editors, *Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data - First China Conference, CCKS 2016, Beijing, China, September 19-22, 2016, Revised Selected Papers*, volume 650 of *Communications in Computer and Information Science*, pages 3–14. Springer, 2016.

10. T. F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome biology*, 6(3):1–8, 2005.
11. L. Huang and X. Luo. EASA: entity alignment algorithm based on semantic aggregation and attribute attention. *IEEE Access*, 8:18162–18170, 2020.
12. J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In K. Chen, C. Huang, and R. Sproat, editors, *Proceedings of the 10th Research on Computational Linguistics International Conference, ROCLING 1997, Taipei, Taiwan, August 1997*, pages 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997.
13. P. Kolyvakis, A. Kalousis, and D. Kiritsis. Deepalignment: Unsupervised ontology matching with refined word vectors. In M. A. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 787–798. Association for Computational Linguistics, 2018.
14. P. Kolyvakis, A. Kalousis, B. Smith, and D. Kiritsis. Biomedical ontology alignment: an approach based on representation learning. *J. Biomedical Semantics*, 9(1):21:1–21:20, 2018.
15. J. J. Lastra-Díaz, J. Goikoetxea, M. A. H. Taieb, A. García-Serrano, M. B. Aouicha, and E. Agirre. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. *Eng. Appl. Artif. Intell.*, 85:645–665, 2019.
16. C. Leacock, M. Chodorow, and G. A. Miller. Using corpus statistics and wordnet relations for sense identification. *Comput. Linguistics*, 24(1):147–165, 1998.
17. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
18. W. Li, X. Duan, M. Wang, X. Zhang, and G. Qi. Multi-view embedding for biomedical ontology matching. In P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, and C. Trojahn, editors, *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 of *CEUR Workshop Proceedings*, pages 13–24. CEUR-WS.org, 2019.
19. Y. Li, Z. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, 15(4):871–882, 2003.
20. D. Lin. An information-theoretic definition of similarity. In J. W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pages 296–304. Morgan Kaufmann, 1998.
21. F. Liu, Y. Shen, T. Zhang, and H. Gao. Entity-related paths modeling for knowledge base completion. *Frontiers Comput. Sci.*, 14(5):145311, 2020.
22. A. N. Ngomo and S. Auer. LIMES - A time-efficient approach for large-scale link discovery on the web of data. In T. Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2312–2317. IJCAI/AAAI, 2011.
23. N. F. Noy, M. A. Musen, J. L. V. M. Jr., and C. Rosse. Pushing the envelope: challenges in a frame-based representation of human anatomy. *Data Knowl. Eng.*, 48(3):335–359, 2004.
24. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.*, 19(1):17–30, 1989.
25. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 448–453. Morgan Kaufmann, 1995.
26. D. Rinser, D. Lange, and F. Naumann. Cross-lingual entity matching and infobox alignment in wikipedia. *Inf. Syst.*, 38(6):887–907, 2013.
27. D. Ritze and H. Paulheim. Towards an automatic parameterization of ontology matching tools based on example mappings. In *Proc. 6th ISWC ontology matching workshop (OM), Bonn (DE)*, pages 37–48, 2011.
28. G. Sogancioglu, H. Öztürk, and A. Özgür. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinform.*, 33(14):i49–i58, 2017.
29. Z. Sun, W. Hu, Q. Zhang, and Y. Qu. Bootstrapping entity alignment with knowledge graph embedding. In J. Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4396–4402. ijcai.org, 2018.
30. B. D. Trisedya, J. Qi, and R. Zhang. Entity alignment between knowledge graphs using attribute embeddings. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 297–304. AAAI Press, 2019.
31. B. D. Trisedya, G. Weikum, J. Qi, and R. Zhang. Neural relation extraction for knowledge base enrichment. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 229–240. Association for Computational Linguistics, 2019.
32. D. Vrandecic. The rise of wikidata. *IEEE Intell. Syst.*, 28(4):90–95, 2013.
33. H. Wang, F. Zhang, X. Xie, and M. Guo. DKN: deep knowledge-aware network for news recommendation. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1835–1844. ACM, 2018.

34. Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao. Relation-aware entity alignment for heterogeneous knowledge graphs. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5278–5284. ijcai.org, 2019.
35. Z. Wu and M. S. Palmer. Verb semantics and lexical selection. In J. Pustejovsky, editor, *32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings*, pages 133–138. Morgan Kaufmann Publishers / ACL, 1994.
36. K. Yang, S. Liu, J. Zhao, Y. Wang, and B. Xie. COTSAE: co-training of structure and attribute embeddings for entity alignment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3025–3032. AAAI Press, 2020.
37. G. Zhu and C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Trans. Knowl. Data Eng.*, 29(1):72–85, 2017.
38. Y. Zhuang, G. Li, Z. Zhong, and J. Feng. Hike: A hybrid human-machine method for entity alignment in large-scale knowledge bases. In E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, and C. Li, editors, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1917–1926. ACM, 2017.