Journal of Data Intelligence, Vol. 3, No. 3 (2022) 316–332

MATCHING LARGE BIOMEDICAL ONTOLOGIES USING SYMBOLIC REGRESSION[a]

JORGE MARTINEZ-GIL
Software Competence Center Hagenberg GmbH
Softwarepark 32a, 4232 Hagenberg, Austria
jorge.martinez-gil@scch.at

SHAOYI YIN
IRIT Toulouse
Route de Narbonne 118, 31062 Toulouse Cedex, France
shaoyi.yin@irit.fr

JOSEF KÜNG
Johannes Kepler University Linz
Altenbergerstraße 69, 4040 Linz, Austria
josef.kueng@jku.at

FRANCK MORVAN
IRIT Toulouse
Route de Narbonne 118, 31062 Toulouse Cedex, France
franck.morvan@irit.fr

The problem of ontology matching consists of finding the semantic correspondences between two ontologies that, although belonging to the same domain, have been developed separately. Ontology matching methods are of great importance today since they allow us to find the pivot points from which an automatic data integration process can be established. Unlike the most recent developments based on deep learning, this study presents our research efforts on the development of novel methods for ontology matching that are accurate and interpretable at the same time. For this purpose, we rely on a symbolic regression model (implemented via genetic programming) that has been specifically trained to find the mathematical expression that can solve the ground truth provided by experts accurately. Moreover, our approach offers the possibility of being understood by a human operator and helping the processor to consume as little energy as possible. The experimental evaluation results that we have achieved using several benchmark datasets seem to show that our approach could be promising.

Keywords: Information Integration, Ontology Matching, Large Ontologies, Semantic Similarity Measures

[a]This paper is an extended version of: Jorge Martinez-Gil, Shaoyi Yin, Josef Küng, Franck Morvan: Matching Large Biomedical Ontologies Using Symbolic Regression. iiWAS 2021: 162-167

1. Introduction

Ontology matching is a field for finding semantic correspondences between ontologies belonging to the same domain but developed separately. Despite its importance in many computer-related disciplines, several problems are currently associated with systems for automatically matching ontologies. For example, the existing matching systems do not allow the discovery of complex correspondences or the fact that most of the existing semantic similarity measures to discover similar entities across ontologies cannot be aggregated easily.

In recent years, it has been an explosion in the number of new techniques and tools for ontology matching aiming to fill these gaps to overcome these problems. These techniques and tools have been a leap in quality compared to the state-of-the-art [1, 2] because they have solved many issues related to accuracy, recall, aggregation, speed of computation, etc. However, there are still some open issues to solve the problem almost definitively. In this paper, we address one of these open issues: interpretability, i.e., the potential ability of a human operator to understand a matching model that has been derived analytically using some computational learning technique.

Biomedical ontology matching, sometimes also called biomedical ontology alignment, consists of finding the semantic correspondences between entities belonging to two ontologies from the biomedical domain developed independently by different experts. One of the main characteristics of this domain is that biomedical ontologies are usually considered large when most state-of-the-art approaches are merely applicable for small-scale ontologies. This usually means that the effectiveness of the existing approaches decreases for large ontologies. This makes our challenge slightly different from the usual one, matching many small ontologies (called holistic matching). For example, it is not very efficient to use structural information since it provides little information concerning such a large model. Furthermore, there is an additional problem when determining these approaches because reference correspondences are unknown in advance, so domain experts must assess samples of the mappings proposed and returned results.

Due to the decentralized nature of biomedical research, the problem is that there usually exist multiple ontologies from overlapped application domains or even within the same domain. In order to establish interoperability between biomedical applications that use different but related ontologies, ontology matching has been proposed as an effective way of handling the semantic heterogeneity problem [3]. It is typically valuable for some data management applications, such as information integration, merging, and distributed query processing. Some ontology matching techniques based on machine learning have recently obtained remarkable results in the biomedical domain. However, the problem is that machine learning methods rely heavily on the availability of high-quality labeled data.

Moreover, if we look exclusively at the latest computational techniques based on deep learning (DL), we find another problem: the solutions from this field often behave like black boxes that users find difficult to understand and trust. The reason is that solutions based on deep neural networks can accept input and provide an output but often do not allow the human operator to understand what happened inside the model before arriving at that output, since there are hundreds if not thousands of nodes and connections between them. This is a severely limiting factor that hinders the development of novel methods and tools for ontology alignment in the biomedical field.

Today, many biomedical applications require matching large biomedical ontologies, so we have focused on addressing this challenge using highly interpretable matching methods based on symbolic regression (via genetic programming). In this way, the significant contributions of this work can be summarized in the following way:

1. The present study extends our previous work on the matching of very large ontologies within the biomedical domain [4]. So that,
    - We propose a method to automatically match large biomedical ontologies based on the concept of symbolic regression intended to facilitate the interpretability of the resulting matching models.
    - We empirically evaluate this new method using popular benchmark datasets in the biomedical domain and offer a comparison with the most prominent ontology matching tools.
2. We emphasize the importance of interpretability and obtaining interpretable models in the biomedical domain today.
3. We bring to the table new results in which the models obtained are analyzed in detail, including a performance analysis.

The rest of this work is structured as follows: Section 2 describes the state-of-the-art regarding ontology matching in the biomedical field. We extend our previous work [4] by deepening aspects related to program synthesis and how this approach is able to improve the interpretability of the models. Section 3 presents our technical contribution including examples in the use of symbolic regression for building aggregation functions. Section 4 extends previous work by describing the benchmark datasets in this biomedical domain, presents the results we have achieved after performing several experiments, and compares our results with other prominent approaches. Please note that we extend our previous work [4] with a convergence analysis for our solution. Finally, we remark on the strengths and weaknesses of our proposal and discuss the future work.

2. State-of-the-art

We review below some of the most significant works in the field of biomedical ontology matching, as well as key aspects for solving the issue of semantic interoperability in the biomedical domain and the importance that program synthesis for symbolic regression has played in recent times in a number of proposed solutions.

2.1. *Ontology matching and semantic interoperability*

One of the earliest studies to formalize this problem was [5], which addressed how ontology matching systems may be created through a trade-off between precision and recall. These authors were the first to propose that instead of using a matching algorithm, machine learning models should be used to generate the most efficient and effective ensemble of matchers.

From this seminal work, the study turned to the proposal of machine learning methods to aggregate the fundamental matchers, with GAOM [6] and GOAL [7, 8] being the first studies

being able to construct the ensemble using genetic algorithms. The fundamental notion was that it might optimize the precision, recall, or combination of the two, for example a harmonic mean called f-measure. In [9], a survey on ontology matching was presented with emphasis on the aggregation methods used by the different proposals existing to date, as well as the significant differences between the techniques for matcher combination, matcher self-tuning, and meta-matching.

There is a large body of literature on methods and tools to address the problem of ontology matching. For example, using high-quality semantic similarity measures [10]. From the classical Latent Semantic Analysis techniques, which prevailed for decades [11], to the most innovative techniques based on word embeddings [12], through the dictionary-based techniques that yielded the best results during the early and mid-2000s [13, 14]. Today, DL is able to obtain the best results [15]. However, these methods are not so good in situations where the interpretability of the solution also plays an important role.

It is important to remark that dealing with large ontologies is a problem that entails a higher complexity than usual. The many homonyms and relationships that only apply in narrow subject domains will lead to many incorrect matches. For general use cases, methods based on embeddings can yield better results. Nowadays, one of the most well-known approaches using embeddings is that of Kolyvakis et al. [16]. The authors propose that ontological term vectors based on information derived from textual corpora and other resources be used to solve the problem. Wu et al. have also lately used Siamese networks to outperform the state-of-the-art in specific instances [17]. Both approaches, however, are based on DL. This implies that they have significant interpretability issues.

In order to overcome this problem, there is work that relies on the use of fuzzy logics; specifically Mamdani models [18], to build solutions that can have some meaning for the users who use them. In particular, fuzzy systems based on Mamdani's postulates are quite interpretable since they operate on the basis of simple rules. However, these techniques still require some specialization and mastery of fuzzy logics, which is not necessarily the case for granted among domain experts from most of applied disciplines.

2.2. *Semantic interoperability in the biomedical domain*

The biological field is one of the domains where the effective and efficient use of ontologies and other knowledge models can significantly impact. Biomedical ontologies are a rich source of information that can help developers create applications for biomedical data annotation, knowledge discovery, decision making, data interoperability, and so on. In this scenario, mappings between the entities corresponding to each ontology are critical for interoperability between data sources. However, heterogeneity is a critical issue which makes the task of finding semantic correspondences difficult. To overcome this problem, several solutions have been already proposed [19, 20, 21, 18, 22, 23, 24].

Regarding the input models to be used, some biomedical ontologies such as the National Cancer Institute Thesaurus (NCI) [25], the SNOMED ontology [26], and the Foundational Model of Anatomy (FMA) [27] to determine the quality of new solutions in this context. Most solutions try to work with as many features as possible. These features include terminological, structural, extensional (instances of a given concept), and external resources. Therefore, the quality of the results has been restricted to small scenarios [28].

There have been several attempts to establish the biomedical ontology alignment challenge as a binary classification problem, i.e., a classifier could be trained with a sample of positive and negative examples provided by that user to identify the cases once it is put into production correctly. However, the results cannot yet be considered optimal. In recent times, DL-based solutions seem to have succeeded in overcoming almost all traditional limitations, including outstanding performance in terms of accuracy in a wide range of scenarios [16, 17]. This is due to a large number of homonyms and associations that only apply to specific subject domains, which will result in a large number of false matches. It is widely assumed that embedding-based matching will produce better results in most applications. However, there are still some gaps, such as the interpretability of the resulting model to be addressed. We address this problem below.

2.3. *Symbolic Regression via Program Synthesis*

Program synthesis is a machine learning technique that searches the space of all programs to find the model that best fits a given dataset. To date, this line of research has been little explored in the field of ontology matching. However, we believe that it possesses a number of characteristics that make it a promising line of research that can bring significant results to the table. We are interested in a particular case of program synthesis where we perform symbolic regression by aggregating ontology matchers of high accuracy through genetic programming, a technique that makes the models evolve under a scheme similar to the evolution of living beings by trying to mimic some cross-over, selection and mutation operations.

As is the case with neural networks, this technique is used to discover solutions to problems people do not know how to address, but unlike networks of a neural nature, the model presents high degrees of interpretability since a person with basic knowledge can understand it. The reason is that the resulting model is expressed in mathematical language and can be easily understood by a person with basic notions of mathematics.

This work presents a symbolic mathematical expression to identify the relationship between defined input and output variables. The use of mathematical expressions allows for great flexibility. In our case, the output is the ontology matching score associated with a pair of entities, while the input variables are values coming from highly interpretable basic matching algorithms already proposed in the literature. In this way, the search space of candidate expressions is enormous. Therefore, it is a much more difficult task than other kinds of regression, such as linear or polynomial regression. The great advantage is that the learned model is a mathematical equation that can be examined and interpreted in the given situation. That resulting equation is chosen to fit the input data and explain the resulting model's functional explanation.

3. Matching Ontologies Using Symbolic Regression

One of the most well-known applications of semantic technologies in general, and ontologies in particular, is the domain of life sciences. Ontologies are models representing regulated terminologies that allow people and machines to understand the meaning of data legibly. In this way, one of the primary goals of biomedical ontologies is to express classes of items relevant to the ontology's development context, which almost always consists of a complex reality full of details and nuances. However, in addition to the names associated with these

classes, the relationships between the various classes are also meaningful. Let us see some definitions of our approach.

3.1. *Definitions and problem formulation*

In the following, we formulate some necessary definitions to follow our contribution. These definitions belong to the domain of semantic similarity measurement since this is the branch of research that we are considering in this work.

Definition 1 (Similarity Function). *A similarity function sf is a function* $sf : \mu_1 \times \mu_2 \mapsto R$ *that associates the similarity of two input pieces of information* μ_1 *and* μ_2 *to a similarity score* $sc \in \Re$ *in the range [0, 1].*

So that a score of 0 stands for absolute inequality and 1 for equality of the pieces of information μ_1 and μ_2 being compared.

Definition 2 (Ontology Matching). *An ontology matching om is a function* $om : O_1 \times O_2 \overset{sm}{\to} A$ *that associates two input ontologies* O_1 *and* O_2 *to an alignment A using a similarity function sf.*

Definition 3 (Ontology Alignment). *An ontology alignment oa is a set* $\{t, MD\}$*, whereby t is a set of tuples in the form* $\{(id, e, e', n, R)\}$*, being id a unique identifier, e and* e' *are entities belonging to two different ontologies, R is the relation of correspondence between these entities, and n is a real number between 0 and 1 that representing the plausibility that R may be true. The entities that can be related are the classes or the relationships of the ontologies. Furthermore, MD is some metadata related to the process for statistical purposes.*

Definition 4 (Alignment Evaluation). *An alignment evaluation ae is a function* $ae : A \times A_R \mapsto precision \in \Re \in [0, 1] \times recall \in \Re \in [0, 1]$ *that associates an alignment A and a reference* A_R *to two real numbers stating the precision, recall of A in relation to* A_R*.*

Definition 5 (Meta-Matching Function). *A Meta-Matching Function* mmf *is a function* $mmf : SC \mapsto \Re$ *that defines how previously calculated similarity score* $sc_i \in SC$*. The result is an optimized similarity score* $sc_o \in \Re$. We call optimized similarity score to the best possible similarity score.

In our case, the meta-matching function will be built using symbolic regression via genetic programming. Genetic programming uses evolutionary strategies to search for one good model from a vast space of solutions representing all the mathematical expressions dealing with the input data. Evolutionary strategies are computational algorithms that work under the assumption of combining two good solutions to create a superior solution. Evolutionary strategies are helpful because they do not use a straightforward optimization approach, allowing for a wide range of outcomes. Furthermore, the resulting model frequently comes up with innovative solutions that provide new insights into the problem.

3.2. *Symbolic Regression*

Symbolic regression is a kind of mathematical analysis in which the model that best fits a given input dataset is found by searching the entire space of all conceivable mathematical expressions. Symbolic regression has already been utilized to solve specific function identification problems in the past. This is primarily owing to the concept of Abstract Syntax Tree (AST), which allows identifying the underlying function from previously solved samples.

Furthermore, the resulting model could be immediately exportable in the form of a program into most of programming languages, making it easier for a human operator to comprehend and transfer to other situations of a similar nature, as we have already shown in [29]. This AST can evolve thanks to an underlying evolutionary strategy.

There are several possibilities to implement such an evolution, but at the end, the definitive result can be calculated by evaluating each node and then performing the parent node operation on the child nodes. Figure 1 shows an AST capable of automatically aggregating three QA methods (m1, m2, and m3). This AST can evolve thanks to the genetic algorithm we referred to before, as explained in [30].

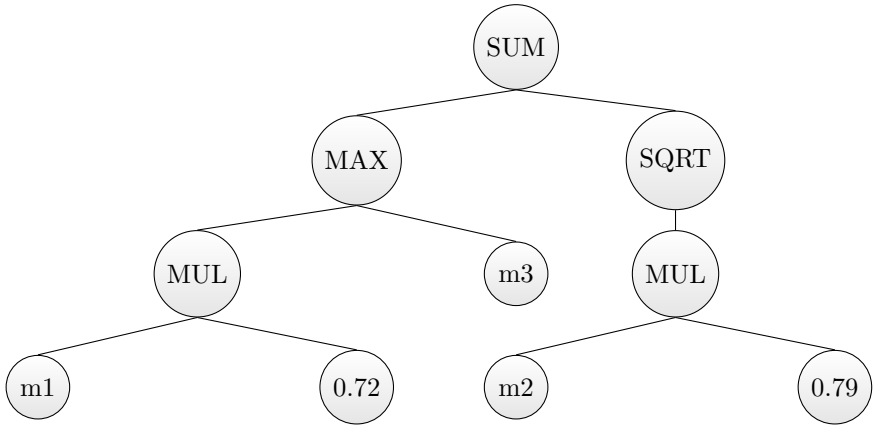

Fig. 1. Example of AST for representing the mathematical expression f(m1, m2, m3) = max((m1*0.72), m3) + sqrt (m2 * 0.79))

Our goal is for AST to evolve to find a mathematical expression that perfectly fits the input-output pairs provided as a training dataset to use that expression over a test dataset. One of the additional advantages of symbolic regression models is that they also allow us to optimize the precedence of operators, which gives even more computing power to the model. Also, not letting the AST grow too much helps us avoid over-fitting problems. This is mainly because simple models behave better in terms of generalization of solutions.

To do that, we aim to aggregate existing matching methods strategically. Aggregation methods are widely used in many disciplines and are often used in production environments, as they allow to avoid the risk of generating completely erroneous results by using highly accurate methods, e.g., [31, 32, 33]. In rare cases where all the methods might simultaneously make the same mistake, the aggregation methods lose their effectiveness. Nevertheless, that is a risk that is not expected to happen often. In any case, this risk will be subjected to an empirical evaluation in the following sections.

3.3. *Role of the intepretability*

The area of eXplainable Artificial Intelligence (XAI) [34] has significantly developed in recent years due to the high interest of both academia and industry for solutions that are not only accurate but can also be understood by the people who handle them [35]. This area has grown so much that in some situations a few hundredths more accuracy does not justify the use of black box models that no one can understand how they work.

In this sense, it can be stated that XAI has a double objective: on the one hand, it seeks to obtain models that are easy to understand but still have high success rates when it comes to acquiring new knowledge. On the other hand, the aim is to make people more capable of understanding how the models they are working with work to act accordingly when appropriate.

In the literature, several terms or expressions are used to refer to this phenomenon, e.g., interpretability, explainability, transparency, accountability, understandability, etc. Each has different nuances. In this work, we focus exclusively on the notion of interpretability, that is the capability of finding a way of representing knowledge that allows the user to fully understand the models to work with.

Currently, there are several models and computational approaches that seek this goal. However, they try to do it in different ways. For example, decision trees attempt to model a flow diagram with decisions acting as steering nodes. Fuzzy logics uses linguistic terms that resemble human language to model and deal with uncertainty. The approach we propose here is based on symbolic regression since we assume that expressing the model in a fully functional form has several advantages concerning its simplicity.

Using other models, such as fuzzy logics, also leads to more interpretable models, as shown in [18]. However, many issues inherent to such models must be studied: their structure, readability, the number of terms involved, the number and length of the rules that implement them, etc. By operating with symbolic regression, we can reduce the problem to a mere question of complexity.

Last but not least, we know that there are three levels of model interpretability: Application-level (only an expert can understand the model), User level (anyone should be able to understand the model), Functional level (the model is expressed as a function). Our approach is the first, to the best of our knowledge, to reach the functional level [36] since anyone with basic mathematical studies can understand it, especially if it is taken into account that we will choose a set of mathematical operators that are very basic.

4. Experimental evaluation

This section presents the empirical study to which we have subjected our approach. We have divided the section into the following subsections: We first describe the nature of the datasets we are working with. Secondly, we explain the metrics used to assess the results obtained. Third, we report the configuration we have used to obtain the results. Fourth, we provide the raw results obtained by our approach and rigorous comparison with state-of-the-art. Fifth, we perform a convergence study of each of the experiments considered in this work. Finally, we describe how our approach can properly model a trade-off between accuracy and interpretability. Furthermore, finally, we discuss the highlights of the whole empirical study.

4.1. *Datasets*

The datasets used in ontology matching are generally based on large-scale open-source data sources. In this work, we have focused on ontology from the biomedical domain as already reported by Kolyvakis et al. [16]. We consider here:

- The foundational Model of Anatomy (FMA) represents the phenotypic structure of the human body [27]. Its goal is to create a conceptual framework of the actual items and regions of the human body.
- The Adult Mouse Anatomical Dictionary (MA) represents the anatomy of an adult mouse [37]. The MA ontology is represented as a graph whose edges represent the relations between the different entities of the mouse.
- The NCI Thesaurus (NCI) represents standard terminology for cancer [25]. We work here with its anatomy subdomain, since it describes occurring human biological structures, and compounds.
- The SNOMED ontology (SNOMED) represents medical nomenclature to be used in clinical reports [26]. It is considered as standard for the electronic exchange of clinical health information in the context of information systems.

Table 1 shows some examples of semantic correspondences between the ontologies we are working with. As can be seen, there are several differentiated cases, from very similar text strings to totally different ones (although with the same semantic content), including those with similar tokens but in a different order.

FMA	**NCI**	**Correspondence**
crown of tooth	corona dentis	True
pericardial artery	pericardiac artery	True
peritoneal fluid	peritoneal effusion	True
pulmonary surfactant	lung surfactant	True
liver parenchyma	hepatic tissue	True
skull bone	human cranium	True

Table 1. Some examples of semantic correspondences between the FMA and NCI ontologies

The number of positive samples is one of the features in this scenario. Because experts typically only provide positive correspondences between ontologies, this is the case. Negative cases are all ones that are not positive; thus, there is no purpose in mentioning them. However, because most regression ensembles presuppose the processing of balanced datasets, this oversupply of positive samples makes the learning process challenging (datasets with a similar number of positive and negative samples). In practice, the number of identical entities between two biomedical ontologies is many orders of magnitude lower than the total number of conceivable combinations.

In addition, it should be noted that there are many correspondences between the biomedical ontologies that are almost similar as we have already seen. This is very common in daily tasks, whereby everybody uses the exactly lexicography.

4.2. *Evaluation criteria*

To evaluate the results of our empirical study, we will use the classical criteria of an information retrieval problem, as do most studies in this context. For this, we will use the traditional metrics based on precision, recall, and f-measure. In the context of ontology matching, precision indicates the the fraction of retrieved mappings (or semantic correspondences) that are relevant to the scenario. The recall is the fraction of the relevant mappings (or semantic correspondences) that are successfully retrieved. F-measure combines precision and recall is the harmonic mean of precision and recall.

In the literature, it is widely assumed that accuracy can be optimized at the expense of recall and vice versa. For this reason, it is convenient to report the two measures together. The f-measure is also interesting, although it cannot model deviations from the metrics on which it is based.

4.3. *Empirical study*

To assess the performance of our approach concerning the state-of-the-art, we also analyzed some of the best proposals in this context. This list is based on the compiled list made by [17]. In fact, we collect most of existing solutions and all the variants based on DL, in addition to the following solutions: AML [19], DOME [22], FCAMapKG [24], LogMapBio[20], and POMAP++ [21].

Table 2 shows the results for the MA-NCI that contains 1489 positives from 9 million possible correspondences. Please note that the ground truth for this experiment is based on the work of Bodenreider et al. [38].

Method	**precision**	**recall**	**f-measure**	**Interpretability**
OM-TD (TF-IDF)	0.900	0.648	0.785	No
OM-TD (LSTM)	0.968	0.704	0.815	No
OM-TD (TBERT)	0.977	0.702	0.817	No
OM (LSTM + SGAT)	0.975	0.717	0.826	No
OM (TBERT + GraphSAGE)	0.954	0.529	0.681	No
OM (TBERT + TransE)	0.890	0.502	0.642	No
OM (DAEOM)	0.981	0.748	0.849	No
DOME	0.993	0.615	0.760	No
AML	0.950	0.936	0.943	Application Level
FCAMapKG	0.996	0.631	0.772	Application Level
LogMapBio	0.872	0.925	0.898	Application Level
POMAP++	0.919	0.877	0.897	Application Level
Our approach	**0.962**	**0.873**	**0.916**	**Functional Level**

Table 2. Results obtained for the MA-NCI

Table 3 shows the results for the benchmark FMA-NCI. The ground truth for this experiment is based on the UMLS Metathesaurus [39] and it contains 2504 positive cases from 24 million possible correspondences.

Table 4 shows the results for the FMA-SNOMED benchmark dataset. Once again, the ground truth for this experiment is based on the UMLS Metathesaurus [39]. It has 7774 positives from 136 million possible correspondences.

Method	precision	recall	f-measure	Interpretability
OM-TD (TF-IDF)	0.969	0.734	0.835	No
OM-TD (LSTM)	0.958	0.871	0.912	No
OM-TD (TBERT)	0.966	0.878	0.920	No
OM (LSTM + SGAT)	0.971	0.879	0.923	No
OM (TBERT + GraphSAGE)	0.981	0.738	0.843	No
OM (TBERT + TransE)	0.961	0.558	0.706	No
OM (DAEOM)	0.989	0.888	0.936	No
DOME	0.985	0.764	0.861	No
AML	0.958	0.910	0.933	Application Level
FCAMapKG	0.967	0.817	0.886	Application Level
LogMapBio	0.919	0.912	0.915	Application Level
POMAP++	0.979	0.814	0.889	Application Level
Our approach	**0.907**	**0.848**	**0.878**	**Functional Level**

Table 3. Results obtained for the FMA-NCI

Method	precision	recall	f-measure	Interpretability
OM-TD (TF-IDF)	0.941	0.613	0.742	No
OM-TD (LSTM)	0.972	0.687	0.805	No
OM-TD (TBERT)	0.977	0.715	0.826	No
OM (LSTM + SGAT)	0.981	0.732	0.838	No
OM (TBERT + GraphSAGE)	0.913	0.677	0.777	No
OM (TBERT + TransE)	0.722	0.506	0.595	No
OM (DAEOM)	0.990	0.791	0.879	No
DOME	0.988	0.198	0.330	No
AML	0.923	0.762	0.835	Application Level
FCAMapKG	0.973	0.222	0.362	Application Level
LogMapBio	0.931	0.703	0.801	Application Level
POMAP++	0.906	0.260	0.404	Application Level
Our approach	**0.907**	**0.770**	**0.832**	**Functional Level**

Table 4. Results obtained for the FMA-SNOMED

4.4. *Convergence Analysis*

We present an analysis of the convergence when training our solution. To do that, we look for the Pareto non-dominated solution front that solves the problem in the best way, i.e. maximizes the f-measure, and how it evolves along the time. Figure 2a depicts the training that was carried out in order to maximize the f-measure of the solution for the MA-NCI challenge.

Because we are using stochastic methods, the results shown are an average of 20 independent experiments from which we depict the worst case (red), the median case (black), and the best case (blue).

Figure 2b depicts the evolutionary process it took to correctly set up our technique to solve the FMA-NCI challenge. As in the preceding scenario, the plotted results are the outcome of 20 independent experiments in which we show again the minimum, median, and maximum values obtained. Figure 2c shows how convergence has occurred when working on the FMA-SNOMED experiment. The plots have the same characteristics as the previous ones.

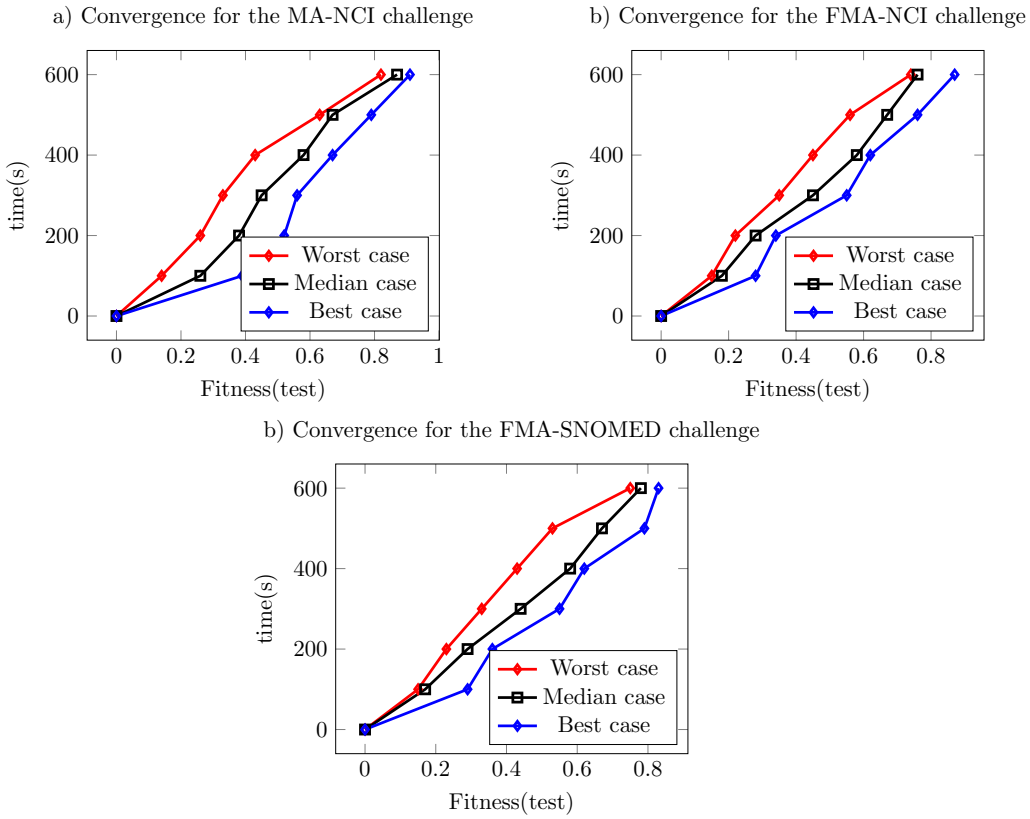

Fig. 2. Fitness evolution for the various experiments that have been carried out. The three plots represent how the f-measure improves over time. The red, black, and blue colors represent the worst, median and best cases respectively

4.5. *Modeling trade-off between accuracy and interpretability*

As a demonstration of what is possible in our solution but not possible to do with other approaches, we model a trade-off between accuracy and interpretability. Since both are considered orthogonal objectives, we can formulate the problem as a bi-objective optimization. In such a problem, we look for the Pareto non-dominated solution front that solves the problem in the best way (e.g., maximum f-measure and interpretability simultaneously).

In our specific case, interpretability is given by a smaller number of items in the resulting AST (and thus the equation). By non-dominated solutions, we mean solutions that can no longer improve one of the two objectives except at the expense of the other. All the solution fronts have been obtained with the multi-objective NSGA-II [40] algorithm which is considered a reference in the field of optimization.

Figure 3a shows us the Pareto front of non-dominated solutions obtained when solving the MA-NCI challenge using our approach. Figure 3b shows us the Pareto front of non-dominated solutions obtained when solving the FMA-NCI challenge using symbolic regression. Last, but not least, Figure 3c shows us the Pareto front of non-dominated solutions obtained when solving the FMA-SNOMED challenge using our proposed technique.

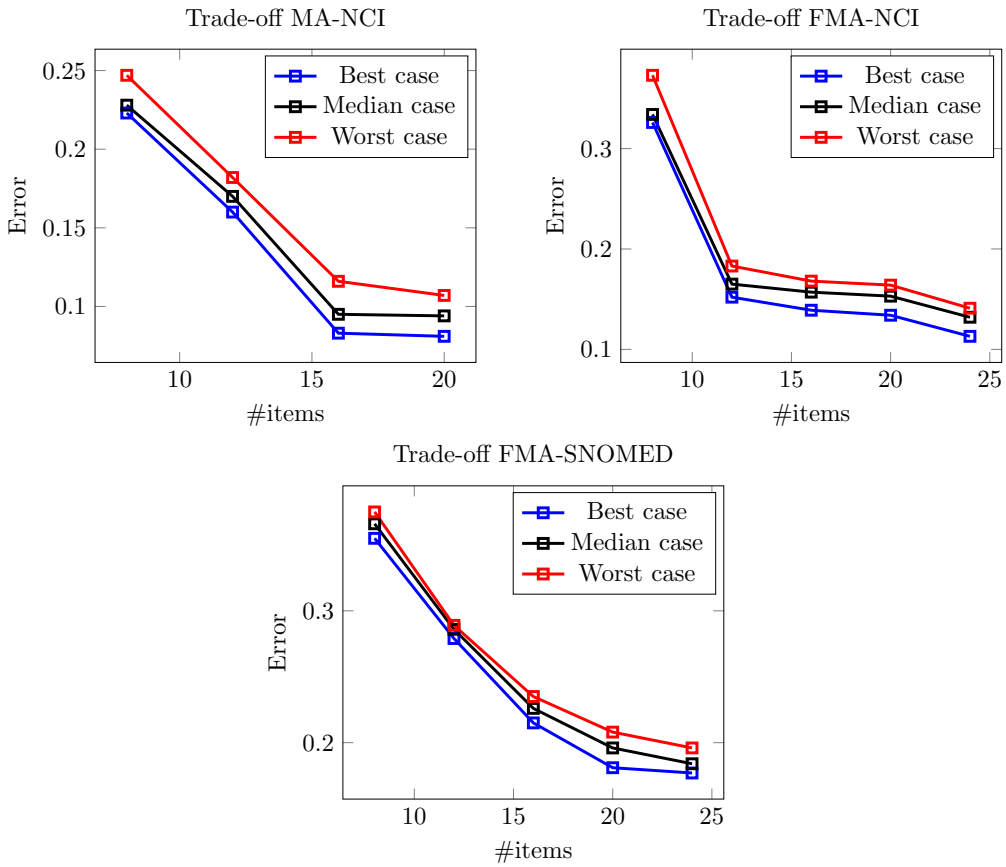

Fig. 3. a) Pareto front of non-dominated points obtained when solving the MA-NCI dataset, b) Pareto front of non-dominated points obtained when solving the FMA-NCI dataset, c) Pareto front of non-dominated points obtained when solving the FMA-SNOMED dataset

These experiments make it possible to appreciate some degree of novelty of our solution compared to existing approaches to represent a trade-off between accuracy and interpretability. This means that human operators can have a set of options in front of them at all times, allowing them to choose the configuration that best matches the problem at hand. This is the first time (if we exclude our previous work [4]) this option has been made available to the human operator in this domain. The reason for this is that, whereas matching approaches have provided users with a trade-off between precision and recall, we are unaware of any existing methodology to represent a trade-off between f-measure and interpretability.

5. Conclusions and Future Work

In this work, we have seen how multiple biomedical ontologies are often created for different applications for the same knowledge domain. The starting point for creating a biomedical ontology is diverse, and the developed ontologies also vary in coverage, granularity, naming characteristics, and structure. As result, these ontologies are highly heterogeneous and bring problems of interoperability.

In order to make full use of the knowledge of different ontological models, it is usually a good idea to integrate them. An essential step for integration is to match the entities in other ontologies that refer to the same real entity.

We have also seen that, when computational methods for big data and data analysis are essential players in biomedical research, the need for people to trust the data-driven systems they use for their daily operations is crucial. However, in recent times, the field of ontology matching, particularly biomedical ontology matching, has been involved in a race to improve accuracy over and over again. This issue has caused little attention to the interpretability of the increasingly accurate solutions.

We have also shown, when it comes to matching biomedical ontologies, our novel method based on the concept of symbolic regression is able to generate acceptable results from both the accuracy and the interpretability perspectives. Moreover, new matcher combinations can be tested based on the results. Even though only a few have been chosen, our method allows us to aggregate any matcher without degrading the overall performance. This way, it may simplify biological data sharing and integration across heterogeneous sources and considerably accelerate biomedical data repositories applications.

Future work, aside from looking for ways to improve accuracy, interpretability, and model good mixes of both measures, also needs to design novel solutions tailored to users' preferences. It is widely assumed that more research in this area is required before the idea of replicating the human operator when dealing with problems related to the automatic matching of biomedical ontologies becomes a reality. Furthermore, regarding program synthesis, although there is literature around the synthesis of code capable of solving operating problems such as classification or regression the attempt to encode background knowledge to further bias the synthesis process has not been explored much yet. We think it may be a promising direction because it can combine at the same time the accuracy of symbolic regression with the expertise and knowledge provided by domain experts from the biomedical field.

Acknowledgements

The authors thank the anonymous reviewers for their help in improving the work. This research has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Ministry of Science, Research and Economy, and the State of Upper Austria through the COMET center SCCH, and by the project FR06/2020 from International Cooperation & Mobility of the Austrian Agency for International Cooperation in Education and Research (OeAD). We would also thank 'the French Ministry of Foreign and European Affairs' and 'The French Ministry of Higher Education and Research' which support the Amadeus program 2020 (French-Austrian Hubert Curien Partnership – PHC) Project Number 44086TD.

References

1. Cássia Trojahn dos Santos. *Towards ontology matching maturity: contributions to complex, holistic and foundational ontology matching.* 2019.
2. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition.* Springer, 2013.
3. Luke T. Slater, Georgios V. Gkoutos, and Robert Hoehndorf. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. *BMC Medical Informatics Decis. Mak.*, 20-S(10):311, 2020.

4. Jorge Martinez-Gil, Shaoyi Yin, Josef Küng, and Franck Morvan. Matching large biomedical ontologies using symbolic regression. In Eric Pardede, Maria Indrawan-Santiago, Pari Delir Haghighi, Matthias Steinbauer, Ismail Khalil, and Gabriele Kotsis, editors, *iiWAS2021: The 23rd International Conference on Information Integration and Web Intelligence, Linz, Austria, 29 November 2021 - 1 December 2021*, pages 162–167. ACM, 2021.
5. Yoonkyong Lee, Mayssam Sayyadian, AnHai Doan, and Arnon Rosenthal. etuner: tuning schema matching software using synthetic scenarios. *VLDB J.*, 16(1):97–122, 2007.
6. Junli Wang, Zhijun Ding, and Changjun Jiang. GAOM: genetic algorithm based ontology matching. In *Proceedings of The 1st IEEE Asia-Pacific Services Computing Conference, APSCC 2006, December 12-15, 2006, Guangzhou, China*, pages 617–620. IEEE Computer Society, 2006.
7. Jorge Martinez-Gil, Enrique Alba, and José Francisco Aldana-Montes. Optimizing ontology alignments by using genetic algorithms. In Christophe Guéret, Pascal Hitzler, and Stefan Schlobach, editors, *Proceedings of the First International Workshop on Nature Inspired Reasoning for the Semantic Web, Karlsruhe, Germany, October 27, 2008*, volume 419 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
8. Jorge Martinez-Gil and José Francisco Aldana-Montes. Evaluation of two heuristic approaches to solve the ontology meta-matching problem. *Knowl. Inf. Syst.*, 26(2):225–247, 2011.
9. Jorge Martinez-Gil and José Francisco Aldana-Montes. An overview of current ontology meta-matching solutions. *Knowl. Eng. Rev.*, 27(4):393–412, 2012.
10. Xingsi Xue and Jie Zhang. Matching large-scale biomedical ontologies with central concept based partitioning algorithm and adaptive compact evolutionary algorithm. *Appl. Soft Comput.*, 106:107343, 2021.
11. Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
12. Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko 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.
13. Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc_ebiquity-core: Semantic textual similarity systems. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*, pages 44–52. Association for Computational Linguistics, 2013.
14. Yuhua Li, Zuhair Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, 15(4):871–882, 2003.
15. Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. Biomedical ontology alignment with BERT. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021*, volume 3063 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org, 2021.
16. Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. Biomedical ontology alignment: an approach based on representation learning. *J. Biomedical Semantics*, 9(1):21:1–21:20, 2018.
17. Jifang Wu, Jianghua Lv, Haoming Guo, and Shilong Ma. Daeom: A deep attentional embedding approach for biomedical ontology matching. *Applied Sciences*, 10(21):7909, 2020.
18. Jorge Martinez-Gil and Jose Manuel Chaves-Gonzalez. Interpretable ontology meta-matching in the biomedical domain using mamdani fuzzy inference. *Expert Syst. Appl.*, 188:116025, 2022.
19. Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F. Cruz, and Francisco M. Couto. The agreementmakerlight ontology matching system. In Robert Meersman, Hervé Panetto, Tharam S. Dillon, Johann Eder, Zohra Bellahsene, Norbert Ritter, Pieter De Leenheer, and Dejing Dou, editors, *On the Move to Meaningful Internet Systems: OTM 2013 Conferences - Confeder-*

ated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings, volume 8185 of *Lecture Notes in Computer Science*, pages 527–541. Springer, 2013.

20. Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. Logmap: Logic-based and scalable ontology matching. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, volume 7031 of *Lecture Notes in Computer Science*, pages 273–288. Springer, 2011.
21. Amir Laadhar, Faiza Ghozzi, Imen Megdiche, Franck Ravat, Olivier Teste, and Faïez Gargouri. Pomap++ results for OAEI 2019: Fully automated machine learning approach for ontology matching. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia 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 169–174. CEUR-WS.org, 2019.
22. Dominique Ritze and Heiko Paulheim. Towards an automatic parameterization of ontology matching tools based on example mappings. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.
23. Xingsi Xue, Chao Jiang, Jie Zhang, and Cong Hu. Biomedical ontology matching through attention-based bidirectional long short-term memory network. *J. Database Manag.*, 32(4):14–27, 2021.
24. Mengyi Zhao, Songmao Zhang, Weizhuo Li, and Guowei Chen. Matching biomedical ontologies based on formal concept analysis. *J. Biomed. Semant.*, 9(1):11:1–11:27, 2018.
25. Sherri de Coronado, Margaret W. Haber, Nicholas Sioutos, Mark S. Tuttle, and Lawrence W. Wright. NCI thesaurus: Using science-based terminology to integrate cancer research results. In Marius Fieschi, Enrico W. Coiera, and Jack Yu-Chan 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.
26. Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.
27. Natalya Fridman Noy, Mark A. Musen, José L. V. Mejino Jr., and Cornelius Rosse. Pushing the envelope: challenges in a frame-based representation of human anatomy. *Data Knowl. Eng.*, 48(3):335–359, 2004.
28. Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. Deepalignment: Unsupervised ontology matching with refined word vectors. In Marilyn A. Walker, Heng Ji, and Amanda 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.
29. Jorge Martinez-Gil and Jose M. Chaves-Gonzalez. A novel method based on symbolic regression for interpretable semantic similarity measurement. *Expert Syst. Appl.*, 160:113663, 2020.
30. John R. Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.
31. Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
32. John W Ratcliff and David E Metzener. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.
33. Gizem Soganciogl u, Hakime Öztürk, and Arzucan Özgür. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinform.*, 33(14):i49–i58, 2017.

34. Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.*, 296:103473, 2021.
35. Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In Karolj Skala, Marko Koricic, Tihana Galinac Grbac, Marina Cicin-Sain, Vlado Sruk, Slobodan Ribaric, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Predrag Pale, and Matej Janjic, editors, *41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, Opatija, Croatia, May 21-25, 2018*, pages 210–215. IEEE, 2018.
36. Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
37. Terry F Hayamizu, Mary Mangan, John P Corradi, James A Kadin, and Martin Ringwald. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome biology*, 6(3):1–8, 2005.
38. Olivier Bodenreider, Terry F. Hayamizu, Martin Ringwald, Sherri de Coronado, and Songmao Zhang. Of mice and men: Aligning mouse and human anatomies. In *AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005*. AMIA, 2005.
39. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270, 2004.
40. Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197, 2002.