GeoLink Cruises: A Non-Synthetic Benchmark for Co-Reference Resolution on Knowledge Graphs

Reihaneh Amini
Kansas State University
reihanea@ksu.edu

Lu Zhou
Kansas State University
luzhou@ksu.edu

Pascal Hitzler
Kansas State University
hitzler@ksu.edu

ABSTRACT

Since over a decade coreference resolution systems have been developed in order to find simple 1-to-1 equivalent mapping (sameAs relations) between instances of different linked datasets and knowledge graphs. Comparative evaluations of instance matching systems can inform us about the performance of such systems regarding artificial benchmarks or real-world data challenges. However, the lack of real data for evaluating these systems is currently a bottleneck. In this paper, we propose the use of the Cruise entities in the GeoLink data repository as a real-world instance matching benchmark for linked data and knowledge graphs. The GeoLink project has brought together seven datasets related to geoscience research. Both the ontology (T-box) and the instance data (A-box) of GeoLink are significantly larger than current benchmarks, and they have particularly interesting challenges, such as geospatial and temporal data. The benchmark we propose here consists of two real-world datasets in GeoLink called R2R data and BCO-DMO which includes manual curated owl:sameAs links between more than 900 Cruise entities of these two datasets. The reference alignment was discussed and generated by domain experts from different institutions and is expressed in the Alignment API format.

CCS CONCEPTS

• **Information systems → Information integration**; **Resource Description Framework (RDF)**; **Web Ontology Language (OWL)**; **Ontologies**; • **Computing methodologies → Ontology engineering**.

KEYWORDS

Coreference Resolution; Instance Matching; Semantic Web; Benchmarks; Linked Data; Knowledge Graphs

ACM Reference Format:
Reihaneh Amini, Lu Zhou, and Pascal Hitzler. 2020. GeoLink Cruises: A Non-Synthetic Benchmark for Co-Reference Resolution on Knowledge Graphs. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412770

1 INTRODUCTION

A massive amount of data is being piped into the world every second by different sources but there is currently not a consistent and steady approach to regulate the format and forms of this data. Linked data emerged to address challenges in data integration and its goal was to introduce techniques and standards to make data accessible and consumable by both humans and machines. According to Tim Berners-Lee, the Semantic Web is not concerned only with putting data on the web but is also about establishing links between that data in order to put it in context. Who should generate these links, though? Often, related data is published by different people at different times, and the publishers may have less knowledge of one another. As a result, there is a need to create systems capable of identifying meaningful links between datasets in an automated way. This is the goal of coreference resolution (alternatively called instance matching). Coreference resolution systems attempt to determine when two URIs refer to the same individual. Unlike complete ontology alignment systems, which endeavor to find many types of relationships (e.g. equivalence, subsumption) between schema-level entities, coreference resolution systems generally focus on identifying 1-to-1 sameAs relationships between instances.

Because the strong performance of coreference resolution systems is vital to harnessing the full power of linked data, it is important to have a benchmark with real data and real challenges with which to evaluate their performance and to spur innovation in the field. In this paper, we propose the GeoLink Cruises benchmark, based on two different real-world geoscience research datasets that were integrated as part of the NSF GeoLink project. We show that this benchmark is relevant to assessing the performance of existing coreference resolution systems in linked data, while also possessing features with the potential to drive innovation in the field.

The main contributions of this paper are therefore the following:

- Presentation of two ontologies to support semantic- and schema-based evaluations.
- Presenting a benchmark with curated sameAs links between over 900 Cruise instances.
- Publication of the benchmark dataset and alignment under a CC-BY license.

This paper is organized as follows. Section 2 discusses the current coreference resolution benchmarks in the Ontology Alignment Evaluation Initiative (OAEI). In Section 3, we provide information about the GeoLink knowledge graph including some of its specific features. Section 4 discusses the coreference resolution benchmark that we are proposing, followed by some analysis of the performance of current coreference resolution algorithms on our data in Section 5. In Section 6, we conclude with a discussion of potential future work in this area.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '20, October 19–23, 2020, Virtual Event, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412770

2 RELATED WORK

In this section, we provide an overview of current main benchmarks in evaluating coreference resolution systems in linked semantic data and also discuss different algorithms for coreference resolution systems and use that as a basis for enumerating a set of desirable qualities for a benchmark with which to evaluate the performance of such systems.

2.1 Coreference Resolution Benchmarks

There are several existing coreference resolution benchmarks in semantically linked data. The presiding existing benchmark is the Ontology Alignment Evaluation Initiative (OAEI) [1], which has held an instance matching track since 2009 [8]. The OAEI has been the foremost venue for researchers focused on ontology alignment and coreference resolution to evaluate and showcase their work. The OAEI introduced this track in 2009, and every year since then, a few coreference resolution systems have participated in the track. This track evaluates system performance by comparing the degree of similarity between pairs of instances with a synthetic benchmark called the Semantic Publishing Instance Matching Benchmark (SPIMBENCH) [18]. SPIMBENCH allows for systematic scalability testing, supports a wider range of test cases, and provides an enriched gold standard. The goal of SPIMBENCH track in OAEI is to resolve coreferences to the same Creative Work. The schema of SPIMBENCH contains 22 classes, 85 object properties, and 31 data properties. SPIMBENCH creates datasets by making various transformations to the source file. Value transformations involve syntactic changes such as misspellings, abbreviations, and synonyms. Structural changes involve modifying the depth of the representation, e.g. by nesting or aggregating properties. Semantic variations involve OWL constructs such as introducing equivalence, disjointness, or complex class definitions between schema level entities. So all these modifications to the source data will generate a set of matches and non-matches as synthetic test cases. So SPIMBENCH does not contain any coreferences within itself because of its synthetic type. SPIMBENCH has also two versions: a small-scale “sandbox” and a medium-scale “mainbox”. Because the data is syntactically generated, the scalability is a modifiable factor and if scalability wants to be tested, SPIMBENCH can synthetically generate many more test cases to evaluate the coreference resolution system’s scalability aspect.

SPIMBENCH is very valuable because complete reference alignments allow both precision and recall to be assessed. In addition, the performance of coreference resolution systems in the face of specific types of characteristics, such as semantic or structural variation, can be isolated and evaluated by the synthetic value tasks. On the other hand, these tasks are of necessity somewhat narrowly focused, and it is not clear how predictive performance on these tasks will be on coreference resolution tasks that differ significantly in scope or domain. In particular, both the source and target dataset use the same ontology (because the datasets contain information from the same domain) which also is not similar to real-world data. Having a semantic-aware aspect is a powerful side of SPIMBENCH benchmark because, in the real semantic world, data includes the RDFS axioms. However, ontologies may include more OWL axioms that should be included in the benchmarks to evaluate the performance of the systems regarding a wide and unexpected range of these axioms, not only a limited set of them. Examples of such axioms include owl:topDataProperty, owl:bottomDataProperty, owl:thing, owl:nothing, owl:real, and so on. Although a synthetic benchmark is very helpful for evaluating systems on many narrow areas, it ingests bias into the evaluation because the false test cases are built on assumptions like string similarity between equivalent classes, and real-world cases may, and do, look different. For example, “paint” and “color” are instances of the same class without any string similarity, or in another case, “Organizer” might refer to the class of “Person” in one dataset and the class of “Company” in another dataset.

Other synthetic instance matching benchmarks independent of the OAEI are similar. For instance, ONTOBI generates ontologies with a moderate number of instances from Wikipedia infoboxes. The size of the reference ontology is again small (17 classes, 13 object properties, and 128 data properties) and the source and target datasets cover the identical domain [22].

As concluding the above-mentioned points, having synthetic data is valuable in evaluating the scalability of the systems and also to avoid the messiness of real data in many cases. In particular, having a synthetic dataset is very important for initial research; however, it cannot fully replicate real-world cases and examples. As a result, there should be both synthetic datasets for initial research and real-world datasets for replicating the complexity of the real data and evaluating the performance of IM systems facing the real world.

The GeoLink Cruise benchmark differs in that both its schema and A-box are considerably more complex since the data is real and the two datasets are based on different larger schemas.

2.2 Coreference Resolution Systems

There have been a total of 20 unique instance matching systems in semantic structured data. We have reviewed all of these systems in order to determine the techniques that are common to many different systems. The results of this activity show that current systems can be classified into a few general approaches, as described below.

Direct comparison of all individual pairs: These approaches often perform surprisingly well but are not scalable to large datasets. Examples include STRIM [13], SLINT++ [15], and ObjectCoref [11]. Systems in this category sometimes do some analysis before beginning the matching task to determine what information to use to compare individuals. For example, SLINT+ chooses properties with good coverage and strong discriminating power while ObjectCoref uses sameAs and inverseFunctional properties to create a training set for learning weightings for property-value pairs.

Comparison of individuals based on a “cloud”: In these approaches, a collection of values is created for each individual, and two individuals are compared using a set-similarity metric over their collections. What is included in the collections varies from system to system? For example, InsMT+ [13] includes the individual’s label and the names of properties specified for that individual. SBUEI [19] includes the individual’s label, its datatype property values, and the datatype values of its direct neighbors. Anchor-Flood

[1] http://oaei.ontologymatching.org/2019/

[9] is similar to SBUEI except that it includes the property names in addition to their values, and only for the individual in question rather than its neighbors.

Two phase comparison: coarse- and fine-grained: These algorithms avoid expensive comparisons between each individual in the source and target ontologies by using faster but less accurate comparisons to find candidate matches that are then compared using more expensive techniques. Some approaches for the coarse-grained comparisons are finding all target individuals that have a data property value exactly in common with the source individual (EXONA [5] and Serimi [1]), finding all individuals that share a property (RiMOM [23]), and finding all individuals of the same type with a somewhat similar label (AgreementMaker [4]) or an exact match on label or alias (Zhishi.links) [16]. The fine-grained comparisons then find the best match from among all of the candidates based on either label (EXONA, RiMOM, AgreementMaker, and Serimi) or property names and values (Zhishi.links).

Reformulation as a different type of problem: Several systems reformulate the coreference resolution problem as a different type of problem. For example, Lily [21] creates a subgraph for each individual based on an electrical circuit model, CODI [12] transforms the alignment problem into the maximum a posteriori optimization problem using Markov logic, and LN2R [17] treats establishing similarities between individuals as a system of equations.

Based on this review of current systems, we conclude that a coreference resolution benchmark should meet the following criteria:

- A large number of instances, to test scalability and approaches that involve coarse-grained filtering.
- Sufficiently numerous common properties among instances of a given type to support comparison on information beyond instance labels. Sufficiently numerous distinct properties among instances of a given type to test systems that combine schema and instance alignment.
- A rich enough schema to support structural-based approaches.
- Enough sameAs links between instances to support approaches that involve supervised machine learning.

Considering the features in the current benchmark and the points regarding what benchmarks need for properly evaluating these systems, we will discuss how our benchmark is addressing these drawbacks.

3 GEOLINK ONTOLOGY MODELS

The benchmark we propose in this paper is a part of the GeoLink knowledge graph. In this section, we explain what the GeoLink knowledge graph is and how our benchmark emerges from it. GeoLink was a project as part of the NSF EarthCube initiative and included seven diverse geoscience datasets that have been merged into a single data repository. GeoLink's merged ontology is available online[2], and the merged data is accessible through SPARQL queries[3]. There are currently 282 classes, 338 properties, 5,118,150 instances, and 45,093,750 triples in the knowledge base.

The overall goal of GeoLink was to merge and unify geoscience data from existing repositories into a single knowledge graph. GeoLink has two different ontologies provided to data providers for publishing their data as RDF triples. These two schemas are the GeoLink Modular Ontology (GMO) and the GeoLink Base Ontology (GBO) [14], which we discuss briefly in this section.

The GeoLink ontologies are developed from ontology design patterns (ODPs) [10] to the high standard of a modular ontology [3, 14]. Each concept like a person or a physical sample that is very frequent in many geoscience repositories was defined as an encapsulated concept (a *module*) in GeoLink in collaboration with a group of domain experts, ontologists, and data providers. All these modules stitched together to form the GMO which enables data provides to publish the part of their data that is related to these concepts, based on GMO vocabulary.

The ODPs represent the concepts within GeoLink that unite the different data repositories. Data providers can publish the parts of their data related to these core concepts according to the vocabulary of the GMO. Any elements in a repository that are not related to the GMO are published in any external vocabulary that best fits the data. Although GMO is following best practices in schema modeling, it is hard to always find the exact relation between your data to some of the concepts designed as ODPs [10]. For example, in the GMO there is an AgentRole class that reifies some relationships. To simplify the mapping, the GBO has been developed, it simplifies the patterns in the GMO [3]. Although the GBO itself does not follow high standards of ontology design, it is more convenient for data providers to publish their data to the GBO. Both ontologies are available online[4] under a CC-BY License. The two ontologies in the GeoLink project have also been used to establish a complex ontology alignment benchmark, which has been utilized in OAEI 2018 and 2019 [24, 25]. The complex alignment benchmark consists of the GMO and GBO ontologies and captures the complex relationships between these two ontologies.

Many of the data repositories contributing to the GeoLink knowledge graph are funded by government organizations such as the NSF. the GeoLink knowledge graph data repositories are:

- R2R (Rolling Deck to Repository) [5]
- BCO-DMO (Biological and Chemical Oceanography Data Management Office) [6]
- DataONE (Data Observation Network for Earth) [7]
- IEDA (Interdisciplinary Earth Data Alliance) [8]
- IODP (International Ocean Discovery Program) [9]
- LTER (Long Term Ecological Research Network) [10]
- MBLWHOI (Marine Biological Laboratory Woods Hole Oceanographic Institution) [11]

In our proposed benchmark, we only use two of these repositories called R2R and BCO-DMO. These two datasets are the main repositories for oceanographic cruises in GeoLink.

[2] http://schema.geolink.org
[3] http://data.geolink.org

[4] https://figshare.com/articles/Complex_Alignment_Benchmark_GeoLink_Dataset/5907172
[5] http://www.rvdata.us
[6] http://www.bco-dmo.org
[7] https://www.dataone.org/
[8] http://www.iedadata.org/
[9] http://www.iodp.org/
[10] http://www.lternet.edu/
[11] http://www.mblwhoilibrary.org/

R2R is the Rolling Deck to Repository, collected by U.S. academic research fleet and includes environmental sensor data. This repository collects data from cruise reports, navigation tracks for cruises, funding awards, vessels, expeditions, and so on.

BCO-DMO is the Biological and Chemical Oceanography Data Management Office (BCO-DMO) which keeps the generated data from oceanographic research and publishes this data online. Much of this data is collected during Cruises for which information is also recorded in R2R.

For generating this GeoLink Cruise benchmark, we considered the part of the GMO and GBO which represent a good and deep enough information about the Cruises, because GeoLink has a big ontology and involving the whole GBO and GMO ontologies may distract our focus on Cruise instances. In fact, we considered the Cruise class since oceanographic cruises are central to the repositories and thus are at the heart of the schema, and most of the other classes, object properties, and data properties are reachable via a few edge traversals from Cruise entities. Figures 1 and 2 provide partial schema diagrams for the GBO and GMO ontologies as relevant for our benchmark.

The **Cruise** pattern is an example of a complex pattern [14] and its notion is central to oceanography data. In both ontologies, there are different classes and we explain some that might require further clarification. **Vessel** represents vessels on which cruises are carried out. **Funding Award/Award** pattern describes the funding awards that fund all kinds of ocean science research activities. A **Program** is a collection of things, including cruises, funding awards, activities, and events. **Event** describes generic events, which may include cruises. **Agent** is a generic class that represents an agent (e.g., a person or an organization) and that agent may perform a role as **AgentRole**. **Place** pattern captures spatial information in the Event pattern above and the rest of the ontology. **Organization** describes organizations, including academic institutions, funding agencies, vessel owners, etc. **Person** appears in a variety of contexts such as Chief Scientist on a cruise, Principal Investigator on a project, a participant in a meeting, or creator of a dataset or paper [2].

4 GEOLINK CRUISE CO-REFERENCE RESOLUTION BENCHMARK

Cruise is one of the three main class types (Cruise, People, Organization) in the GeoLink knowledge base. Therefore, we propose an instance matching benchmark called GeoLink Cruise Benchmark including two Cruise datasets, which are R2R and BCO-DMO (see Section 3). The GeoLink Cruise benchmark has some unique features which distinguish it from other current benchmarks. We explain these features in the following.

Accessibility: The benchmark is publicly available[12] under a CC-BY 4.0 License, which means it is free to download, manipulate, and use for any purposes. Both the GMO and GBO directory in the data include the reference alignment between the ontologies named "refalign.rdf". The reference alignment is expressed using the Alignment format[13] by the Alignment API [6], because this is the most prominent syntax for such benchmarks, including SPIMBENCH.

Real-world dataset: The GeoLink project comprises many datasets provided by seven data providers. These datasets have been collected and utilized in the geosciences, which indicates that they are natural, realistic datasets, rather than artificial ones as often used in benchmarks. This is one of the main features of our data which makes it more useful for evaluating the performance of instance matching algorithms for practice. As we discussed earlier in Section 2, assumptions that humans make in regards to generating synthetic data are introducing bias into the data which might lead to biased system evaluation. Having synthetic data is useful for early research and for avoiding the messiness of real data for early system testing. It is also helpful in evaluating the systems in specific areas. However, it cannot sufficiently capture the complexity of real-world data.

Ground Truth: For this benchmark, Cruise instances in BCO-DMO and R2R were compared with each other manually and **491 owl:sameAs** links were generated between them by data providers. This large number of curated links can evaluate the precision and recall of the instance matching systems properly. For evaluating scalability, more data can easily be piped from the GeoLink endpoint for additional assessment.

Dataset Prepreration We tried to bring freedom and flexibility for system evaluation by combining the datasets and ontologies that we have. In this benchmark, there are four different scenarios for selecting two datasets for an evaluation task. We mapped the R2R and BCO-DMO data to both GMO and GBO ontologies which we discussed in Section 3. As a result, we have four datasets in our benchmark:

- **BCO-DMO_GBO** (BCO-DMO data mapped to GBO ontology)
- **BCO-DMO_GMO** (BCO-DMO data mapped to GMO ontology)
- **R2R_GBO** (R2R data mapped to GBO ontology)
- **R2R_GMO** (R2R data mapped to GMO ontology)

Table 1: Base statistics for the GeoLink Cruise Benchmark

Ontology	Classes	Object Properties	Data Properties	Individuals	Triples
BCO-DMO_GBO	40	149	49	1061	13055
R2R_GBO	40	149	49	5320	27992
BCO-DMO_GMO	79	79	37	1052	16303
R2R_GMO	79	79	37	2025	24798

As a result, in the evaluation of coreference resolution systems, any combination of two datasets from these four datasets could provide a different insight into the IM system and measure different capabilities of them. For instance, two datasets with different ontology schema (R2R_GBO and BCO-DMO_GMO) as the source and target data, could be selected for evaluation tasks or two different datasets with the same ontology schema (BCO-DMO_GMO and R2R_GMO) could be picked up.

Table 1 shows the number of classes and properties in all data sources.

Alignment Tasks Our ground truth data includes manually curated equivalence mappings among 491 entities pairs in BCO-DMO and R2R. In coreference resolution tasks, the alignment system will be asked to find these 491 sameAs links. The performance will be evaluated using precision, recall, and f-measure, as the majority of coreference resolution benchmarks also apply these measures to

[12] https://figshare.com/articles/GeoLink_Cruise_Benchmark/12443999

[13] http://alignapi.gforge.inria.fr/

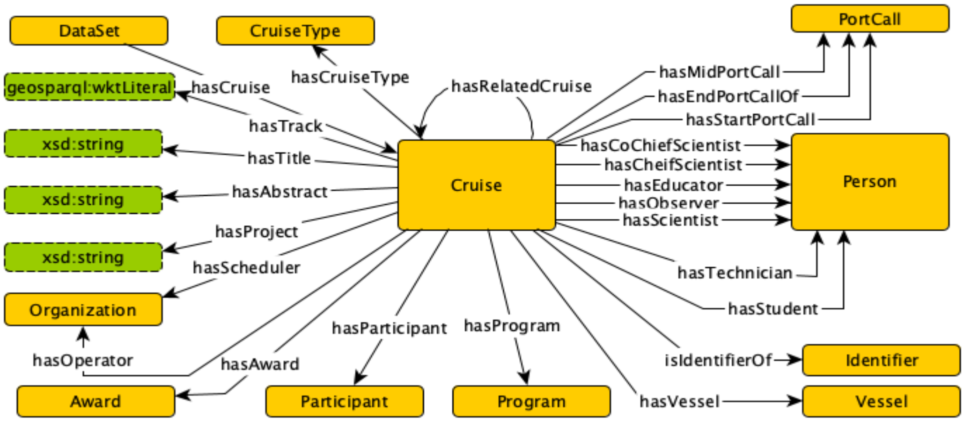

Figure 1: GBO Cruise Ontology Schema

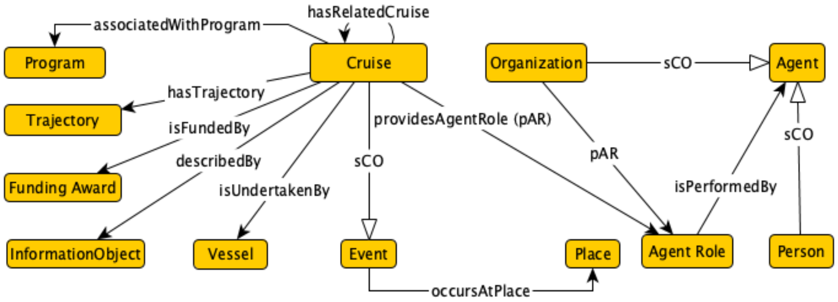

Figure 2: GMO Cruise Ontology Schema, (sCO=subClassOf and pAR=providesAgentRole)

Table 2: owl:sameAs Link Example between Two Cruise Instances in R2R and BCO-DMO.

```
@Prefix bco-dmo: <http://lod.bco-dmo.org/geolink/id/deployment/> .
@Prefix gbo: <https://gbo#> .
@Prefix r2r: <http://data.rvdata.us/id/cruise/> .
@Prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@Prefix foaf: <http://xmlns.com/foaf/0.1/> .
@Prefix rvdata: <http://data.rvdata.us/id/program/> .
```

bco-dmo:57490	gbo:hasProgram	bco-dmo:2012 .
bco-dmo:2012	gbo:hasAbstract	“U.S. GLOBEC (GLOBal ocean ECosystems dynamics)...”;
	gbo:hasTitle	“U.S. GLOBal ocean ECosystems dynamics"@en-us ;
	gbo:hasAcronym	“U.S. GLOBEC” .
r2r:AT7-21	gbo:hasProgram	rvdata:GLOBEC .
rvdata:GLOBEC	rdfs:label	“Global Ocean Ecosystem Dynamics” ;
	gbo:description	“Global Ocean Ecosystem Dynamics” ;
	foaf:name	“Global Ocean Ecosystem Dynamics” .

assess the completeness, soundness, and overall matching quality of the instance matching systems. The alignment tasks between the ontologies are listed as follows:

```
Evaluation between BCO-DMO_GMO and R2R_GMO
Evaluation between BCO-DMO_GBO and R2R_GMO
Evaluation between BCO-DMO_GMO and R2R_GBO
Evaluation between BCO-DMO_GBO and R2R_GBO
```

We expect that coreference resolution systems should perform better in finding the links between the entities of the same schema and semantic-aware comparison should be more straightforward. For example, if we want to evaluate the performance of the IM system in founding the similarities between the datasets with the same ontologies, R2R_GMO and BCO-DMO_GMO could be selected because, in this pair, R2R and BCO-DMO both have GMO ontology. In another combination of the data, comparing R2R_GBO and R2R_GMO is also a good scenario for testing the systems regarding their ability to map identical individuals with different schema.

By providing different datasets with different ontologies, we tried to introduce a benchmark capable of evaluating a system properly in different semantic scenarios.

Ontology Schema: GeoLink has two finely developed schemas designed by domain experts, data providers, and ontologists which provide good support for structure-based approaches in coreference resolution algorithms.

In Figures 1 and 2, it shows all the classes and properties related to Cruise in GBO and GMO respectively. Cruise has such an enriched ontology with various properties (object and data property). For simple string similarities, the property value of some data properties, such as hasTitle, hasProject, and hasAbstract in GBO are easily comparable and for fine-grained comparison, more information could be pulled out by traversing through more object properties and data properties in the graph.

In addition, the GeoLink benchmark includes various OWL and OWL2 axioms that are helpful for evaluating the systems that work based on ontologies. Such axioms utilize OWL vocabulary such as

```
owl:sameAs, owl:priorVersion, owl:versionInfo,
owl:imports, owl:unionOf ,owl:inverseOf,
owl:complementOf.
```

Beyond label comparison: Although current coreference resolution benchmarks with synthetic data provide only value-based test cases through adding typographical errors to data, the GeoLink Cruise benchmark can capture matches that go beyond similar class label and string matches in data properties due to its real nature and size. Table 2 is an example of two cruises in BCO-DMO respectively R2R that are in fact the same cruise. As it is shown, these two cruises have both a property called "hasProgram" and if we further look in the data for data about these programs, there are properties such as "hasTitle", "hasAcronym", "description", "name", and "label", which are pointing to values which are proper strings, making it possible to utilize string similarity metrics. Only systems that are comprehensive enough to look more deeply into the data can capture these kinds of alignments, because the two Cruise instances may not share the same label, but their object properties may share more information. So having real data with challenges beyond data property and label comparison can help to evaluate coreference systems regarding their suitability for realistic data.

5 EXPERIMENTS AND ANALYSIS

In this section, based on the results of the instance matching track at OAEI 2019,[14], we apply AgreementMakerLight (AML) [7] and test it on the proposed benchmark. The AML source code is published on Github,[15] while other systems such as Lily [21] and FTRL-IM [20] are not open source. AML has in fact been actively participating in all the different tracks in OAEI and has achieved excellent performance overall. Therefore, AML can be considered a representative alignment system for the state of the art in the area of instance matching.

As presented in the OAEI result of 2019, AML achieved an f-measure of 86% on SPIMBENCH for both sandbox and mainbox. It performs a proper job of linking objects at both the ontology and instances level. An example of an equivalence link is depicted in Table 3, which AML found between two entities *bbct17:id* and *bbct1261410333:id* in SPIMBENCH. A closer look reveals that these two instances almost share the same triples with some minor changes in one or two properties and property values (changes are in colored text). As we discussed before, the way that SPIMBENCH generates test cases is by adding spelling changes in source data. As you can see here, the string transformation just added some minor changes and most of the data is untouched. For example here the dataProperty "cw:title" and "cw:shortTitle" in the source data is being changed to "cw:title0" and "cw:title1" in the target data in SPIMBENCH. As a result, AML or other systems can easily align the two entities since they don't differ that much from each other considering both property and value. The importance of having real data for evaluation becomes rather obvious from this example and by comparing it with the realistic GeoLink data as displayed, e.g., in Figure 2.

By running AML on our benchmark, the results are different. AML could not generate good coreferences between the instances in the GeoLink benchmark. However, it found some alignments at the ontology level. We show some of the correct and incorrect alignments that AML could find, while it was running on **BCO-DMO_GBO** and **R2R_GMO**.

Example of correct links:

```
<entity1 rdf:resource="http://gbo#hasFamilyName"/>
<entity2 rdf:resource="http://gmo#familyOrSurname"/>

<entity1 rdf:resource="http://gbo#hasAffiliation"/>
<entity2 rdf:resource="http://gmo#hasPosition"/>

<entity1 rdf:resource="http://gbo#isPortOf"/>
<entity2 rdf:resource="http://gmo#r_Port"/>

<entity1 rdf:resource="http://gbo#hasGivenName"/>
<entity2 rdf:resource="http://gmo#firstOrGivenName"/>
```

Example of incorrect links:

```
<entity1 rdf:resource="http://gbo#hasContributor"/>
<entity2 rdf:resource="http://gmo#hasAttribute"/>

<entity1 rdf:resource="http://gbo#hasCurator"/>
```

[14] https://project-hobbit.eu/challenges/om2019/

[15] https://github.com/AgreementMakerLight/AML-Project/releases/tag/v3.1

Table 3: Example of coreference alignment found between two entities of SPIMBENCH by AML algorithm. (colored parts are the transformations that SPIMBENCH made.)

bbct17:id		**bbct1261410333:id**	
@prefix bbct17: http://www.bbc.co.uk/things/17/ @prefix cw: http://www.bbc.co.uk/ontologies/creativework/ @prefix bbct1509989698: <http://www.bbc.co.uk/ things/1509989698/		@prefix bbct1261410333: <http://www.bbc.co.uk/ things/1261410333/ @prefic cw: http://www.bbc.co.uk/ontologies/ creativework/ @prefix bbct1509989698: <http://www.bbc.co.uk/things/ 1509989698/	
rdf:type	cw:Programme	rdf:type	cw:Programme
cw:title	Khalid Mahmood little very provider king until not it accept be or."	cw:title0	Khalid Mahmood little very provide
cw:shortTitle	millions method competitive we wrestling head sometimes participation turn result	cw:title1	r king until not it accept be or.
cw:category	bbcc:PoliticsPersons-Additional	cw:category	http://www.bbc.co.uk/category/ PoliticsPersonsAdditional
cw:description	" craftsmen make maintained pay layout allowing as commerce did politics see." .	cw:description	" craftsmen make maintained pay layout allowing as commerce did politics see."
cw:about	wikidata:Q695028	cw:about	wikidata:Q695028
cw:about	http://dbpedia.org/resource /Cypriot-PresidentialElection2003	cw:about	http://dbpedia.org/resource/ Cypriot-presidentialelection,2003
cw:mentions	http://sws.geonames.org /7299636	cw:mentions	http://sws.geonames.org /7299636/
cw:audience	cw:InternationalAudience	cw:audience	cw:InternationalAudience
cw:liveCoverage	true http://www.w3.org /2001/XMLSchema/boolean	cw:liveCoverage	true http://www.w3.org/2001/ XMLSchema/boolean
cw:primaryFormat	cw:AudioFormat	cw:primaryFormat	cw:AudioFormat
cw:dateCreated	"2011-11-17T20:19:06.076+02:00" http://www.w3.org/2001/ XMLSchema/dateTime	cw:dateCreated	"2011-11-17T20:19:06.076+02:00" http://www.w3.org/2001/ XMLSchema/dateTime
cw:dateModified	"2012-10-10T04:09:55.770+03:00" http://www.w3.org/2001/ XMLSchema/dateTime	cw:dateModified	"2012-10-10T04:09:55.770+03:00" http://www.w3.org/2001/ XMLSchema/dateTime
cw:thumbnail	bbct17:219976210	cw:thumbnail	bbct:219976210
cw:altText	"thumbnail atlText for CW http://www.bbc.co.uk/ context/17/id		
bbc:primaryContentOf	bbct1509989698:id	bbc:primaryContentOf	bbct1509989698:id

```
<entity2 rdf:resource="http://www.w3.org/2006/time
#hasDuration"/>
```

As we have seen, AML found property alignments between GMO and GBO. The system is working fine on finding the links that have a higher string similarity or string synonyms between the two objects, but for the objects with weaker string similarities like "hasAttribute" and "hasContribute", AML usually reported false positives by wrongly adding a mapping between the objects with partial string matching.

In conclusion, having a simple synthetic benchmark might be useful for the evaluation of the systems in specific areas in an early stage. A benchmark comprising real-world data can boost the development of the state of the art alignment algorithms in the field of instance matching.

6 CONCLUSION AND FUTURE WORKS

Coreference resolution on semantically linked data has been discussed for over a decade now, but not much work has been done related to a realistic benchmark for systematic evaluation of these systems. In this paper, we have proposed the use of the GeoLink Cruise data repository as a coreference resolution benchmark. The GeoLink Cruise benchmark utilizes manually curated equivalents links between Cruise instances of R2R and BCO-DMO data. This Benchmark has been accepted by OAEI 2020 and is going to exist for evaluating systems going forward.

An analysis of current automated coreference resolution systems indicates that current synthetic benchmarks are insufficiently replicating the complexities present in real-world data, and our dataset can be used to evaluate entity resolution approaches from a more realistic perspective. Additionally, our dataset has the potential to spur innovations related to scalability and data aligning with geospatial and temporal aspects.

In the future, we plan to actively keep maintaining the benchmark and contribute to its use at the OAEI. In addition, we also plan to assess the performance of several existing approaches to coreference resolution on this benchmark. Furthermore, we intend to develop an alignment system that can meet the challenges proposed by our benchmark.

ACKNOWLEDGMENTS

Lu Zhou and Pascal Hitzler acknowledge support by the National Science Foundation under Grant 1936677 through *Convergence Accelerator Phase I (RAISE): Spatially-Explicit Models, Methods, and Services for Open Knowledge Networks.*

REFERENCES

[1] Samur Araújo, Arjen P. de Vries, and Daniel Schwabe. 2011. SERIMI results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oaei11_paper15.pdf

[2] Michelle Cheatham, Reihaneh Amini, and Chandan Patel. 2016. Matching instances in GeoLink. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016 (CEUR Workshop Proceedings, Vol. 1766)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, Oktie Hassanzadeh, and Ryutaro Ichise (Eds.). CEUR-WS.org, 237–238. http://ceur-ws.org/Vol-1766/om2016_poster6.pdf

[3] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. 2018. The GeoLink knowledge graph. *Big Earth Data* 2, 2 (2018), 131–143.

[4] Isabel F. Cruz, Cosmin Stroe, Federico Caimi, Alessio Fabiani, Catia Pesquita, Francisco M. Couto, and Matteo Palmonari. 2011. Using AgreementMaker to align ontologies for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oaei11_paper1.pdf

[5] Syrine Damak, Hazem Souid, Marouen Kachroudi, and Sami Zghal. 2015. EXONA results for OAEI 2015. In *Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015 (CEUR Workshop Proceedings, Vol. 1545)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, and Oktie Hassanzadeh (Eds.). CEUR-WS.org, 145–149. http://ceur-ws.org/Vol-1545/oaei15_paper5.pdf

[6] Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn dos Santos. 2011. The Alignment API 4.0. *Semantic Web* 2, 1 (2011), 3–10. https://doi.org/10.3233/SW-2011-0028

[7] Daniel Faria, Catia Pesquita, Emanuel Santos, Isabel F. Cruz, and Francisco M. Couto. 2013. AgreementMakerLight results for OAEI 2013. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013 (CEUR Workshop Proceedings, Vol. 1111)*, Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz (Eds.). CEUR-WS.org, 101–108. http://ceur-ws.org/Vol-1111/oaei13_paper1.pdf

[8] Alfio Ferrara, Andriy Nikolov, Jan Noessner, and François Scharffe. 2013. Evaluation of instance matching tools: The experience of OAEI. *J. Web Semant.* 21 (2013), 49–60. https://doi.org/10.1016/j.websem.2013.05.004

[9] Md. Seddiqui Hanif and Masaki Aono. 2009. Anchor-Flood: Results for OAEI 2009. In *Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009 (CEUR Workshop Proceedings, Vol. 551)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Natalya Fridman Noy, and Arnon Rosenthal (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-551/oaei09_paper1.pdf

[10] Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, and Valentina Presutti (Eds.). 2016. *Ontology Engineering with Ontology Design Patterns - Foundations and Applications*. Studies on the Semantic Web, Vol. 25. IOS Press.

[11] Wei Hu, Jianfeng Chen, Gong Cheng, and Yuzhong Qu. 2010. ObjectCoref & Falcon-AO: results for OAEI 2010. In *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010 (CEUR Workshop Proceedings, Vol. 689)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-689/oaei10_paper6.pdf

[12] Jakob Huber, Timo Sztyler, Jan Nößner, and Christian Meilicke. 2011. CODI: Combinatorial Optimization for Data Integration: results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oaei11_paper4.pdf

[13] Abderrahmane Khiat, Moussa Benaissa, and Mohammed Amine Belfedhal. 2015. STRIM results for OAEI 2015 instance matching evaluation. In *Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015 (CEUR Workshop Proceedings, Vol. 1545)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, and Oktie Hassanzadeh (Eds.). CEUR-WS.org, 208–215. http://ceur-ws.org/Vol-1545/oaei15_paper16.pdf

[14] Adila Alfa Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert A. Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Tim Finin, Peng Ji, Matthew B. Jones, Nazifa Karima, Kerstin A. Lehnert, Audrey Mickle, Tom Narock, Margaret O'Brien, Lisa Raymond, Adam Shepherd, Mark Schildhauer, and Peter Wiebe. 2015. The GeoLink Framework for Pattern-based Linked Data Integration. In *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015 (CEUR Workshop Proceedings, Vol. 1486)*, Serena Villata, Jeff Z. Pan, and Mauro Dragoni (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1486/paper_99.pdf

[15] Khai Nguyen and Ryutaro Ichise. 2013. SLINT+ results for OAEI 2013 instance matching. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013 (CEUR Workshop Proceedings, Vol. 1111)*, Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz (Eds.). CEUR-WS.org, 177–183. http://ceur-ws.org/Vol-1111/oaei13_paper11.pdf

[16] Xing Niu, Shu Rong, Yunlong Zhang, and Haofen Wang. 2011. Zhishi.links results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oaei11_paper16.pdf

[17] Fatiha Saïs, Nobal B. Niraula, Nathalie Pernelle, and Marie-Christine Rousset. 2010. LN2R a knowledge based reference reconciliation system: OAEI 2010 results. In *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010 (CEUR Workshop Proceedings, Vol. 689)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-689/oaei10_paper8.pdf

[18] Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. 2015. Pushing the Limits of Instance Matching Systems: A Semantics-Aware Benchmark for Linked Data. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 105–106. https://doi.org/10.1145/2740908.2742729

[19] Aynaz Taheri and Mehrnoush Shamsfard. 2012. SBUEI: results for OAEI 2012. In *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012 (CEUR Workshop Proceedings, Vol. 946)*, Pavel Shvaiko, Jérôme Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Fridman Noy, and Heiner Stuckenschmidt (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-946/oaei12_paper11.pdf

[20] Xiaowen Wang, Yizhi Jiang, Yi Luo, Hongfei Fan, Hua Jiang, Hongming Zhu, and Qin Liu. 2019. FTRLIM Results for OAEI 2019. In *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 (CEUR Workshop Proceedings, Vol. 2536)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn (Eds.). CEUR-WS.org, 146–152. http://ceur-ws.org/Vol-2536/oaei19_paper9.pdf

[21] Jiangheng Wu, Zhe Pan, Ce Zhang, and Peng Wang. 2019. Lily Results for OAEI 2019. In *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 (CEUR Workshop Proceedings, Vol. 2536)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn (Eds.). CEUR-WS.org, 153–159. http://ceur-ws.org/Vol-2536/oaei19_paper10.pdf

[22] Katrin Zaiß, Stefan Conrad, and Sven Vater. 2010. A Benchmark for Testing Instance-based Ontology Matching Methods. In *Proceedings of the EKAW2010 Poster and Demo Track, Lisbon, Portugal, October 11 - 15, 2010 (CEUR Workshop Proceedings, Vol. 674)*, Johanna Völker and Óscar Corcho (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-674/Paper157.pdf

[23] Yan Zhang, Hailong Jin, Liangming Pan, and Juan-Zi Li. 2016. RiMOM results for OAEI 2016. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016 (CEUR Workshop Proceedings, Vol. 1766)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, Oktie Hassanzadeh, and Ryutaro Ichise (Eds.). CEUR-WS.org, 210–216. http://ceur-ws.org/Vol-1766/oaei16_paper13.pdf

[24] Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. 2018. A Complex Alignment Benchmark: GeoLink Dataset. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11137)*, Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl (Eds.). Springer, 273–288. https://doi.org/10.1007/978-3-030-00668-6_17

[25] Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. 2020. GeoLink Data Set: A Complex Alignment Benchmark from Real-world Ontology. *Data Intell.* 2, 3 (2020), 353–378. https://doi.org/10.1162/dint_a_00054