Matching Instances in GeoLink

Michelle Cheatham, Reihaneh Amini, and Chandan Patel

DaSe Lab, Wright State University, Dayton OH 45435, USA,
{michelle.cheatham, amini.2, patel.383}@wright.edu

Abstract. We propose the use of the GeoLink data repository as an instance matching benchmark. The GeoLink project brings together seven datasets related to geoscience research. Both the T-box and the A-box of GeoLink are significantly larger than current benchmarks, and they have interesting challenges, such as geospatial and temporal data.

GeoLink is part of the NSF EarthCube initiative. Seven diverse geoscience datasets have been brought together into a single data repository. The ontology is documented at `http://schema.geolink.org`, and the triple store is accessible at `http://data.geolink.org`. There are currently 282 classes, 338 properties, 5,118,150 instances and 45,093,750 triples in the knowledge base. The are also owl:sameAs and skos:closeMatch links between instances of different types. The sameAs links were manually generated by the data providers, while the closeMatch links were generated by an automated coreference resolution system. We highlight three different classes within the GeoLink schema that pose different opportunities for evaluating and challenging coreference resolution systems: Person, Cruise, and Organization.

Person Instances of Person appear in a variety of contexts such as Chief Scientist on a cruise, Principal Investigator on a project, participant in a meeting, or creator of a dataset or paper. Key object properties related to the person class reflect these different contexts. Related data properties include name, email address, and ORCID.[1] GeoLink considers the NSF dataset to be "canonical" for the Person class, meaning that Person instances in each of the other datasets have been mapped to NSF instances. The NSF dataset contains 335,504 people, so it is not feasible to compare each person from one of the constituent datasets to every person in the NSF datset. This benchmark can therefore be used to encourage development of systems that employ effective filtering or other mechanisms to achieve scalablility. The triple store currently contains 15,660 people not in the NSF dataset. There are 790 sameAs and 1,405 closeMatch links between these people and those within the NSF data.

Cruise There are 12,070 cruises in the GeoLink repository, potentially allowing an m by n comparison. There are 1,356 sameAs links and 368 closeMatch links among cruises. The cruise coreference task is intriguing because cruises have geospatial and temporal elements, which are considered an important challenge

[1] `http://orcid.org`

for coreference resolution systems [3]. Two properties of particular interest are hasTrack and hasPortCall properties. A cruise's track is generally a series of latitude and longitude coordinates. The Cruise class also has properties hasStartPortCall, hasMidPortCall, and hasEndPortCall. The PortCall class is in the domain of the properties hasTimeStamp and hasPort, whereas the range of hasTimeStamp is a date time literal and the range of hasPort is Place. A place can be described in terms of its latitude and longitude, but it might also be identified using a gazetteer term.

Organization Compared to Person and Cruise, the GeoLink knowledge base contains relatively little information about instances of the Organization class. There is often little data other than the organization's title and the set of people who are affiliated with it in the knowledge base. Finding coreferences in this situation is likely to be difficult for approaches that rely on extensive schema information; however, approaches that rely on, for instance, the degree of overlap between the people affiliated with two organizations to measure their similarity, may perform quite well. Because there are nearly 300,000 organizations within GeoLink, this is again a task in which approaches that do not perform some type of filtering are unlikely to be feasible. There are currently no sameAs links between organizations, but 268 closeMatch links have been established.

There are several existing coreference resolution benchmarks. The dominant existing benchmark is that of the OAEI, which has included an instance matching track since 2009 [1]. Some tasks within this track are synthetic (generated via SPIMBENCH [2]) while others are real-world. The benchmark proposed here differs because it is less narrowly focused and involves a much larger schema and A-box. On the other hand, because the current set of links in GeoLink is likely not exhaustive, only recall (and not precision) can be evaluated.

Acknowledgments The authors sincerely thank the GeoLink team.[2] This work was supported by the National Science Foundation GeoLink project (1440202).

References

1. Ferrara, A., Nikolov, A., Noessner, J., Scharffe, F.: Evaluation of instance matching tools: The experience of OAEI. Web semantics: Science, services and agents on the World Wide Web 21, 49–60 (2013)
2. Saveta, T., Daskalaki, E., Flouris, G., Fundulaki, I., Herschel, M., Ngonga Ngomo, A.C.: 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. pp. 105–106. ACM (2015)
3. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176 (2013)

[2] `http://www.geolink.org/team.html`