Mining Equivalent Relations from Linked Data

Ziqi Zhang[1] **Anna Lisa Gentile**[1] **Isabelle Augenstein**[1]
Eva Blomqvist[2] **Fabio Ciravegna**[1]

[1] Department of Computer Science, University of Sheffield, UK
[2] Department of Computer and Information Science, Linköping University, Sweden
{z.zhang, a.l.gentile, i.augenstein, f.ciravegna}@dcs.shef.ac.uk, eva.blomqvist@liu.se

Abstract

Linking heterogeneous resources is a major research challenge in the Semantic Web. This paper studies the task of mining equivalent relations from Linked Data, which was insufficiently addressed before. We introduce an unsupervised method to measure equivalency of relation pairs and cluster equivalent relations. Early experiments have shown encouraging results with an average of 0.75~0.87 precision in predicting relation pair equivalency and 0.78~0.98 precision in relation clustering.

1 Introduction

Linked Data defines best practices for exposing, sharing, and connecting data on the Semantic Web using uniform means such as URIs and RDF. It constitutes the conjunction between the Web and the Semantic Web, balancing the richness of semantics offered by Semantic Web with the easiness of data publishing. For the last few years Linked Open Data has grown to a gigantic knowledge base, which, as of 2013, comprised 31 billion triples in 295 datasets[1].

A major research question concerning Linked Data is linking heterogeneous resources, the fact that publishers may describe analogous information using different vocabulary, or may assign different identifiers to the same referents. Among such work, many study mappings between ontology concepts and data instances (e.g., Isaac et al, 2007; Mi et al., 2009; Le et al., 2010; Duan et al., 2012). An insufficiently addressed problem is linking heterogeneous relations, which is also widely found in data and can cause problems in information retrieval (Fu et al., 2012). Existing work in linking relations typically employ string similarity metrics or semantic similarity measures that require a-priori domain knowledge and are limited in different ways (Zhong et al., 2002; Volz et al., 2009; Han et al., 2011; Zhao and Ichise, 2011; Zhao and Ichise, 2012).

This paper introduces a novel method to discover equivalent groups of relations for Linked Data concepts. It consists of two components: 1) a measure of equivalency between pairs of relations of a concept and 2) a clustering process to group equivalent relations. The method is unsupervised; completely data-driven requiring no a-priori domain knowledge; and also language independent. Two types of experiments have been carried out using two major Linked Data sets: 1) evaluating the precision of predicting equivalency of relation pairs and 2) evaluating the precision of clustering equivalent relations. Preliminary results have shown encouraging results as the method achieves between 0.75~0.85 precision in the first set of experiments while 0.78~0.98 in the latter.

2 Related Work

Research on linking heterogeneous ontological resources mostly addresses mapping classes (or concepts) and instances (Isaac et al, 2007; Mi et al., 2009; Le et al., 2010; Duan et al., 2012; Schopman et al., 2012), typically based on the notions of similarity. This is often evaluated by string similarity (e.g. string edit distance), semantic similarity (Budanitsky and Hirst, 2006), and distributional similarity based on the overlap in data usage (Duan et al., 2012; Schopman et al., 2012). There have been insufficient studies on mapping relations (or properties) across ontologies. Typical methods make use of a combination of string similarity and semantic similarity metrics (Zhong et al., 2002; Volz et al., 2009; Han et al., 2011; Zhao and Ichise, 2012). While string similarity fails to identify equivalent relations if their lexicalizations are distinct, semantic similarity often depends on taxonomic structures

[1] http://lod-cloud.net/state/

in existing ontologies (Budanitsky and Hirst, 2006). Unfortunately many Linked Data instances use relations that are invented arbitrarily or originate in rudimentary ontologies (Parundekar et al., 2012). Distributional similarity has also been used to discover equivalent or similar relations. Mauge et al. (2012) extract product properties from an e-commerce website and align equivalent properties using a supervised maximum entropy classification method. We study linking relations on Linked Data and propose an unsupervised method. Fu et al. (2012) identify similar relations using the overlap of the subjects of two relations and the overlap of their objects. On the contrary, we aim at identifying strictly equivalent relations rather than similarity in general. Additionally, the techniques introduced our work is also related to work on aligning multilingual Wikipedia resources (Adar et al., 2009; Bouma et al., 2009) and semantic relatedness (Budanitsky and Hirst, 2006).

3 Method

Let t denote a 3-tuple (triple) consisting of a subject (t_s), predicate (t_p) and object (t_o). Linked Data resources are *typed* and its type is called *class*. We write *type* (t_s) = c meaning that t_s is of class c. p denotes a relation and r_p is a set of triples whose t_p=p, i.e., r_p=$\{t \mid t_p = p\}$.

Given a specific class c, and its pairs of relations (p, p') such that r_p=$\{t|t_p$=p, *type*(t_s)=$c\}$ and $r_{p'}$=$\{t|t_p$=p', *type* (t_s)=$c\}$, we measure the equivalency of p and p' and then cluster equivalent relations. The equivalency is calculated locally (within same class c) rather than globally (across all classes) because two relations can have identical meaning in specific class context but not necessarily so in general. For example, for the class *Book*, the relations *dbpp:title* and *foaf:name* are used with the same meaning, however for *Actor*, *dbpp:title* is used interchangeably with awards *dbpp:awards* (e.g., Oscar best actor).

In practice, given a class c, our method starts with retrieving all t from a Linked Data set where *type*(t_s)=c, using the universal query language SPARQL with any SPARQL data endpoint. This data is then used to measure equivalency for each pair of relations (Section 3.1). The equivalence scores are then used to group relations in equivalent clusters (Section 3.2).

3.1 Measure of equivalence

The equivalence for each distinct pair of relations depends on three components.

Triple overlap evaluates the degree of overlap[2] in terms of the usage of relations in triples. Let *SO(p)* be the collection of subject-object pairs from r_p and SO_{int} the intersection

$$SO_{int}(p,p') = SO(r_p) \cap SO(r_{p'}) \quad [1]$$

then the triple overlap $TO(p, p')$ is calculated as

$$MAX\{\frac{|SO_{int}(r_p,r_{p'})|}{|r_p|}, \frac{|SO_{int}(r_p,r_{p'})|}{|r_{p'}|}\} \quad [2]$$

Intuitively, if two relations p and p' have a large overlap of subject-object pairs in their data instances, they are likely to have identical meaning. The *MAX* function allows addressing infrequently used, but still equivalent relations (i.e., where the overlap covers most triples of an infrequently used relation but only a very small proportion of a much more frequently used).

Subject agreement While triple overlap looks at the data in general, subject agreement looks at the overlap of subjects of two relations, and the degree to which these subjects have overlapping objects. Let $S(p)$ return the set of subjects of relation p, and $O(p|s)$ returns the set of objects of relation p whose subjects are s, i.e.:

$$O(p \mid s) = O(r_p \mid s) = \{t_o \mid t_p = p, t_s = s\} \quad [3]$$

we define:

$$S_{int}(p,p') = S(r_p) \cap S(r_{p'}) \quad [4]$$

$$\alpha = \frac{\sum_{s \in S_{int}(p,p')} \begin{cases} 1, if \mid O(p \mid s) \cap O(p' \mid s) \mid > 0 \\ 0, \quad otherwise \end{cases}}{|S_{int}(p,p')|} \quad [5]$$

$$\beta = \sqrt{|S_{int}(p,p')| / |S(p) \cup S(p')|} \quad [6]$$

then the agreement $AG(p, p')$ is

$$AG(p,p') = \alpha \cdot \beta \quad [7]$$

Equation [5] counts the number of overlapping subjects whose objects have at least one overlap. The higher the value of α, the more the two relations "agree" in terms of their shared subjects. For each shared subject of p and p' we count 1 if they have at least 1 overlapping object and 0 otherwise. This is because both p and p' can be *1:many* relations and a low overlap value could mean that one is densely populated while the other is not, which does not necessarily mean they do not "agree". Equation [6] evaluates the degree to which two relations share the same set of subjects. The agreement $AG(p, p')$ balances the two factors by taking the product. As a result,

[2] In this paper overlap is based on "exact" match.

relations that have high level of agreement will have more subjects in common, and higher proportion of shared subjects with shared objects.

Cardinality ratio is a ratio between cardinality of the two relations. Cardinality of a relation $CD(p)$ is calculated based on data:

$$CD(p) = \frac{|r_p|}{|S(r_p)|} \quad [8]$$

and the cardinality ratio is calculated as

$$CDR(p, p') = \frac{MIN\{CD(p), CD(p')\}}{MAX\{CD(p), CD(p')\}} \quad [9]$$

The final **equivalency measure** integrates all the three components to return a value in [0, 2]:

$$E(p, p') = \frac{TO(p, p') + AG(p, p')}{CDR(p, p')} \quad [10]$$

The measure will favor two relations that have similar cardinality.

3.2 Clustering

We apply the measure to every pair of relations of a concept, and keep those with a non-zero equivalence score. The goal of clustering is to create groups of equivalent relations based on the pair-wise equivalence scores. We use a simple rule-based agglomerative clustering algorithm for this purpose. First, we rank all relation pairs by their equivalence score, then we keep a pair if (i) its score and (ii) the number of triples covered by each relation are above a certain threshold, $T_{minEqvl}$ and T_{minTP} respectively. Each pair forms an initial cluster. To merge clusters, given an existing cluster c and a new pair (p, p') where either $p \in c$ or $p' \in c$, the pair is added to c if $E(p, p')$ is close (as a fractional number above the threshold $T_{minEqvlRel}$) to the average scores of all connected pairs in c. This preserves the strong connectivity in a cluster. This is repeated until no merge action is taken. Adjusting these thresholds allows balancing between precision and recall.

4 Experiment Design

To our knowledge, there is no publically available gold standard for relation equivalency using Linked Data. We randomly selected 21 concepts (Figure 1) from the DBpedia ontology (v3.8):

Actor, Aircraft, Airline, Airport, Automobile, Band, BasketballPlayer, Book, Bridge, Comedian, Film, Hospital, Magazine, Museum, Restaurant, Scientist, TelevisionShow, TennisPlayer, Theatre, University, Writer

Figure 1. Concepts selected for evaluation.

We apply our method to each concept to discover clusters of equivalent relations, using as *SPARQL endpoint* both DBpedia[3] and Sindice[4] and report results separately. This is to study how the method performs in different conditions: on one hand on a smaller and cleaner dataset (DBpedia); on the other hand on a larger and multi-lingual dataset (Sindice) to also test cross-lingual capability of our method. We chose relatively low *thresholds*, i.e. $T_{minEqvl}$*=0.1,* T_{minTP}*= 0.01%* and $T_{minEqvlRel}$*=0.6,* in order to ensure high recall without sacrificing much precision.

Four human annotators manually annotated the output for each concept. For this preliminary evaluation, we have limited the amount of annotations to a maximum of 100 top scoring pairs of relations per concept, resulting in 16~100 pairs per concept (avg. 40) for DBpedia experiment and 29~100 pairs for Sindice (avg. 91). The annotators were asked to rate each edge in each cluster with -1 (wrong), 1 (correct) or 0 (cannot decide). Pairs with 0 are ignored in the evaluation (about 12% for DBpedia; and 17% for Sindice mainly due to unreadable encoded URLs for certain languages). To evaluate cross-lingual pairs, we asked annotators to use translation tools. Inter-Annotator-Agreement (observed IAA) is shown in Table 1. Also using this data, we derived a gold standard for clustering based on edge connectivity and we evaluate (i) the *precision* of top *n*% (*p@n%*) ranked equivalent relation pairs and (ii) the *precision* of clustering for each concept.

	Mean	High	Low
DBpedia	0.79	0.89	0.72
Sindice	0.75	0.82	0.63

Table 1. IAA on annotating pair equivalency

So far the output of 13 concepts has been annotated. This dataset[5] contains ≈1800 relation pairs and is larger than the one by Fu et al. (2012). Annotation process shows that over 75% of relation pairs in the Sindice experiment contain non-English relations and mostly are cross-lingual. We used this data to report performance, although the method has been applied to all the 21 concepts, and the complete results can be visualized at our demo website link. Some examples are shown in Figure 2.

[3] http://dbpedia.org/sparql
[4] http://sparql.sindice.com/
[5] http://staffwww.dcs.shef.ac.uk/people/Z.Zhang/ resources/paper/acl2013short/web/

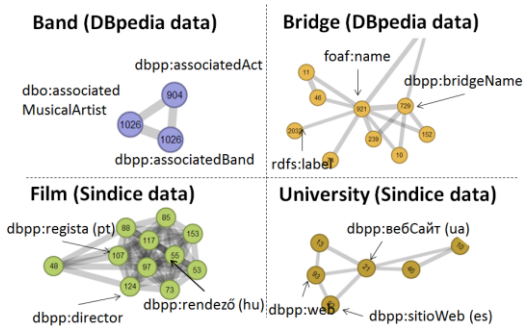

Figure 2. Examples of visualized clusters

5 Result and Discussion

Figure 3 shows *p@n%* for pair equivalency[6] and Figure 4 shows clustering precision.

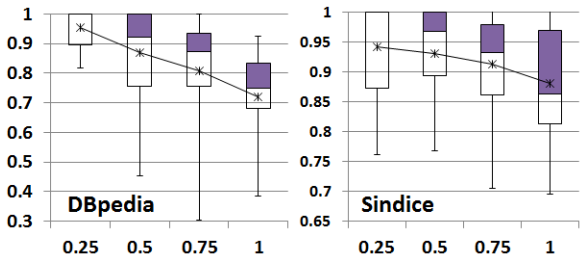

Figure 3. *p@n%*. The box plots show the ranges of precision at each *n%*; the lines show the average.

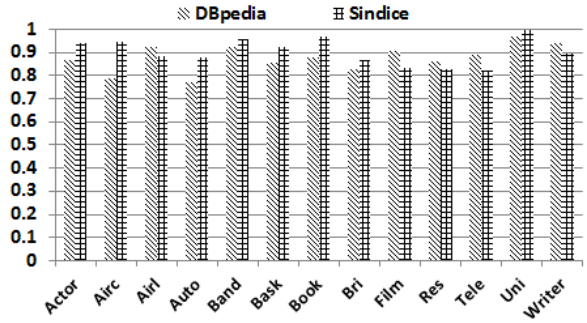

Figure 4. Clustering precision

As it is shown in Figure 2, Linked Data relations are often heterogeneous. Therefore, finding equivalent relations to improve coverage is important. Results in Figure 3 show that in most cases the method identifies equivalent relations with high precision. It is effective for both single- and cross-language relation pairs. The worst performing case for DBpedia is *Aircraft* (for all *n%*), mostly due to duplicating numeric valued objects of different relations (e.g., weight, length, capacity). The decreasing precision with respect to *n%* suggests the measure effectively ranks correct pairs to the top. This is a useful feature from IR point of view. Figure 4 shows that the method effectively clusters equivalent relations with very high precision: 0.8~0.98 in most cases.

[6] Per-concept results are available on our website.

Overall we believe the results of this early proof-of-concept are encouraging. As a concrete example to compare against Fu et al. (2012), for *BasketballPlayer*, our method creates separate clusters for relations meaning "draft team" and "former team" because although they are "similar" they are not "equivalent".

We noticed that annotating equivalent relations is a non-trivial task. Sometimes relations and their corresponding schemata (if any) are poorly documented and it is impossible to understand the meaning of relations (e.g., due to acronyms) and even very difficult to reason based on data. Analyses of the evaluation output show that errors are typically found between highly similar relations, or whose object values are numeric types. In both cases, there is a very high probability of having a high overlap of subject-object pairs between relations. For example, for *Aircraft*, the relations *dbpp:heightIn* and *dbpp:weight* are predicted to be equivalent because many instances have the same numeric value for the properties. Another example are the *Airport* properties *dbpp:runwaySurface, dbpp:r1Surface, dbpp:r2Surface* etc., which according to the data seem to describe the construction material (e.g., concrete, asphalt) of airport runways. The relations are semantically highly similar and the object values have a high overlap. A potential solution to such issues is incorporating ontological knowledge if available. For example, if an ontology defines the two distinct properties of *Airport* without explicitly defining an "equivalence" relation between them, they are unlikely to be equivalent even if the data suggests the opposite.

6 Conclusion

This paper introduced a data-driven, unsupervised and domain and language independent method to learn equivalent relations for Linked Data concepts. Preliminary experiments show encouraging results as it effectively discovers equivalent relations in both single- and multilingual settings. In future, we will revise the equivalence measure and also experiment with clustering algorithms such as (Beeferman et al., 2000). We will also study the contribution of individual components of the measure in such task. Large scale comparative evaluations (incl. recall) are planned and this work will be extended to address other tasks such as ontology mapping and ontology pattern mining (Nuzzolese et al., 2011).

Acknowledgement

Part of this research has been sponsored by the EPSRC funded project LODIE: Linked Open Data for Information Extraction, EP/J019488/1. Additionally, we also thank the reviewers for their valuable comments given for this work.

References

Eytan Adar, Michael Skinner, Daniel Weld. 2009. *Information Arbitrage across Multilingual Wikipedia*. Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 94 – 103.

Gosse Bouma, Sergio Duarte, Zahurul Islam. 2009. Cross-lingual Alignment and Completion of Wikipedia Templates. Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, pp. 61 – 69

Doug Beeferman, Adam Berger. 2000. Agglomerative clustering of a search engine query log. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 407-416.

Alexander Budanitsky and Graeme Hirst. 2006. *Evaluating WordNet-based Measures of Semantic Distance*. Computational Linguistics, 32(1), pp.13-47.

Songyun Duan, Achille Fokoue, Oktie Hassanzadeh, Anastasios Kementsietsidis, Kavitha Srinivas, and Michael J. Ward. 2012. *Instance-Based Matching of Large Ontologies Using Locality-Sensitive Hashing.* ISWC 2012, pp. 46 – 64

Linyun Fu, Haofen Wang, Wei Jin, Yong Yu. 2012. *Towards better understanding and utilizing relations in DBpedia*. Web Intelligence and Agent Systems , Volume 10 (3)

Andrea Nuzzolese, Aldo Gangemi, Valentina Presutti, Paolo Ciancarini. 2011. *Encyclopedic Knowledge Patterns from Wikipedia Links*. Proceedings of the 10th International Semantic Web Conference, pp. 520-536

Lushan Han, Tim Finin and Anupam Joshi. 2011. GoRelations: *An Intuitive Query System for DBpedia*. Proceedings of the Joint International Semantic Technology Conference

Antoine Isaac, Lourens van der Meij, Stefan Schlobach, Shenghui Wang. 2007. *An empirical study of instance-based ontology matching*. Proceedings of the 6th International Semantic Web Conference and the 2nd Asian conference on Asian Semantic Web Conference, pp. 253-266

Ngoc-Thanh Le, Ryutaro Ichise, Hoai-Bac Le. 2010. *Detecting hidden relations in geographic data.* Proceedings of the 4th International Conference on Advances in Semantic Processing, pp. 61 – 68

Karin Mauge, Khash Rohanimanesh, Jean-David Ruvini. 2012. *Structuring E-Commerce Inventory*. Proceedings of ACL2012, pp. 805-814

Jinhua Mi, Huajun Chen, Bin Lu, Tong Yu, Gang Pan. 2009. *Deriving similarity graphs from open linked data on semantic web.* Proceedings of the 10th IEEE International Conference on Information Reuse and Integration, pp. 157–162.

Rahul Parundekar, Craig Knoblock, José Luis. Ambite. 2012. *Discovering Concept Coverings in Ontologies of Linked Data Sources*. Proceedings of ISWC2012, pp. 427–443.

Balthasar Schopman, Shenghui Wang, Antoine Isaac, Stefan Schlobach. 2012. *Instance-Based Ontology Matching by Instance Enrichment*. Journal on Data Semantics, 1(4), pp 219-236

Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov. 2009. *Silk – A Link Discovery Framework for the Web of Data*. Proceedings of the 2nd Workshop on Linked Data on the Web

Lihua Zhao, Ryutaro Ichise. 2011. *Mid-ontology learning from linked data*. Proceedings of the Joint International Semantic Technology Conference, pp. 112 – 127.

Lihua Zhao, Ryutaro Ichise. 2012. *Graph-based ontology analysis in the linked open data*. Proceedings of the 8th International Conference on Semantic Systems, pp. 56 – 63

Jiwei Zhong, Haiping Zhu, Jianming Li and Yong Yu. 2002. *Conceptual Graph Matching for Semantic Search*. The 2002 International Conference on Computational Science.