A Framework to Generate Reference Sets for Ontology Matching Algorithms[*]

Gurpriya Bhatia[0000−0002−7511−8543], Kumar Vidhani[0000−0002−2412−6391], Mangesh Gharote[0000−0002−4942−2429], and Sachin Lodha[0000−0001−5771−4977]

54B, Tata Research Development and Design Center, Tata Consultancy Services Ltd.,
Hadapsar Industrial Estate, Hadapsar, Pune, Maharashtra -411013
{gurpriya.bhatia, kumar.vidhani, mangesh.g, sachin.lodha}@tcs.com

Abstract. The performance of ontology matching algorithms is evaluated using F-measure, precision and recall which in turn rely on the availability of the ground truth. Typically, the ground truth generation process is manual, subjective and time consuming. Therefore, there is a need to come up with a (semi) automated approach which generates an unbiased reference set; an approximation of ground truth. We propose a framework based solution to generate a reference set and report encouraging results for the OAEI 2019 conference dataset.

Keywords: Reference Set · Ontology Matching Algorithm Property · Ontology Matching.

1 Introduction

The performance of ontology matching algorithms is evaluated using the F-measure, precision, and recall measures. These measures in turn rely on the ground truth (gold standard) generated by a community of domain experts. Typically, the ground truth creation is manual, subjective and time consuming exercise. Due to its subjective nature, even the creation of a small size ground truth requires many domain experts to agree on a small set of pairs (e.g., some ontology pairs of the conference data set have less than 15 pairs in their ground truth).

Ground truth is the requirement in almost every scientific discipline to validate ideas, theories, methods, etc. Therefore, many semiautomated approaches are proposed in various domains to generate it. Euzenat et al. propose benchmark generator framework to measure the meaningful properties of ontology matching algorithms [1]. The objective of their framework is to generate a new benchmark by supporting various alteration operations for any seed ontology. DBPediaNYD, another such effort, has resulted in the machine generated reference set (a silver standard)[5]. Jorn Hees has proposed a semiautomated approach to map Edinburgh Associative Thesaurus (EAT) to DBpedia entities [3]. Hees approach finds candidate mappings automatically through scores assigned to them by using Wikipedia API. These mappings are further verified manually to generate final set of mappings. Harrow et al. have evaluated 11 matching systems on the biomedical ontologies to evaluate their relative performance with respect manually created mappings (gold standard), a set of mappings generated through consensus (silver standard or a reference set), and unique mappings generated by individual participating system [2].

Existing approaches do not consider a way to address bias introduced in the reference set as a result of using particular approach to generate it. For example, an algorithm that uses web search engines may get unfair advantage in an evaluation when using DBPedia-NYD as the reference set [5]. Creation of an unbiased reference set offers multiple advantages: i) it can be used to evaluate a newly proposed ontology matching algorithm, ii) it can be used for training purpose, and iii) it can serve as the starting point for generating the ground truth.

2 Framework

We propose a plug-and-play framework that exploits properties of different ontology matching algorithms to generate an unbiased reference set for the input ontology matching algorithm and a pair of ontologies. Figure 1 outlines a conceptual view of the proposed framework. The framework enables the right set of ontology matching algorithms depending on the requirements specified by the user (domain expert, ontology matching algorithm designer, etc). For example, if the user wants to generate a reference set to be used for evaluating an ontology matching algorithm that exploits *distance* property between concepts of input ontologies, the framework enables those ontology matching algorithms which exploit different properties (e.g., *concept equivalence* through synonym set) to avoid bias in the reference set. Further, the user may choose to compute confidence values for all or a subset of concept pairs of input ontologies.

To generate a reference set of desired size and quality, it is necessary to filter the alignment set with respect to threshold values on the size and confidence values computed by all framework algorithms. Algorithm 1 outlines the approach to select threshold on the confidence values for an ontology pair. The selection of

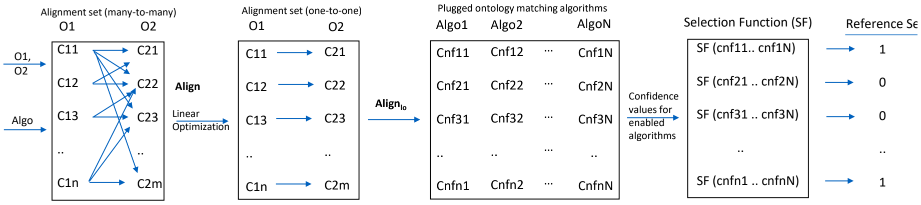

Fig. 1. Conceptual View of Framework.

threshold value τ is determined by two parameters, the cardinality of a set in one-to-one matching form (generated after applying linear optimization - $|algoSet|$ - as shown in algorithm) and $\rho \in [0, 1]$, the user defined parameter for the minimum size of reference set.

Selection Function (SF) is one of the most important elements of the framework. SF takes 'n' confidence values computed by chosen 'n' ontology matching algorithms for a concept pair and produces a boolean value. To put it formally, $SF : [0, 1]^n \rightarrow \{1, 0\}$. Different implementations of the SF function are possible. In its current avatar of the framework, we provide two implementations. First implementation uses *Unanimity rule* approach. All chosen algorithms should agree on a concept pair for its inclusion in the reference set. Second implementation uses *Majority rule* approach. If the majority of ontology matching algorithms ($>= 50\%$) agree on a concept pair, it is included in the reference set.

Algorithm 1 Algorithm to compute threshold value

```
Require: $Algo_{set}$, a superset containing one-to-one matching sets of all framework
    algorithms for an ontology pair, $\rho$, user defined parameter
for all $threshold\ in\ [0.1, .., 1.0]$ do
    $flag = true$
    for all $algoSet \in Algo_{set}$ do
        $filteredSet = filterForThreshold(threshold, algoSet)$
        if $(|filteredSet|/|algoSet|) < \rho$ then
            $flag = false$
        end if
    end for
    if $flag == true$ then
        $setThresholdForOntoPair(threshold)$
    end if
end for
```

3 Experiments

We have conducted experiments on the OAEI 2019 conference dataset using python v3.7.3. We have evaluated our framework using six different ontology matching algorithms two each for the categories of Deep learning (word2vec[1] and fastText[2]), WordNet (WuPalmer and Lin[3]) and character (nGram and MLCS[4]).

For the computation of equality relation, classes and properties are compared with classes and properties respectively. Moreover, we first convert the output of each ontology matching algorithm that is in many-to-many form (*Align*) into

[1] https://spacy.io/api/doc/

[2] https://fasttext.cc/docs/en/pretrained-vectors.html

[3] https://www.nltk.org/howto/wordnet.html

[4] https://pypi.org/project/strsim/

Table 1. Comparison of two implementations of Selection Function for the conference dataset. $\mathrm{F}_{\mathrm{EXDL}}$, $\mathrm{F}_{\mathrm{EXWN}}$ and $\mathrm{F}_{\mathrm{EXCHR}}$ - F-measure values (in percentage) excluding DL, WordNet and Character based approaches respectively. $\rho = 0.1$.

Ontology Pair	Threshold(τ)	SF-*Unanimity rule*			SF-*Majority rule*		
		$\mathrm{F}_{\mathrm{EXDL}}$	$\mathrm{F}_{\mathrm{EXWN}}$	$\mathrm{F}_{\mathrm{EXCHR}}$	$\mathrm{F}_{\mathrm{EXDL}}$	$\mathrm{F}_{\mathrm{EXWN}}$	$\mathrm{F}_{\mathrm{EXCHR}}$
cmt_Conference	0.8	40.00	46.15	40.00	41.17	45.16	44.44
cmt_confOf	0.8	43.47	50.00	43.47	48.00	48.00	46.15
cmt_edas	0.8	63.63	60.86	57.14	61.53	64.00	59.25
cmt_ekaw	0.8	52.63	52.63	60.00	42.85	54.54	40.00
cmt_iasted	0.8	66.66	88.88	75.00	42.10	72.72	44.44
cmt_sigkdd	0.8	70.00	72.72	70.00	69.56	75.00	72.00
Conference_confOf	0.8	58.33	66.66	56.00	48.64	58.06	47.36
Conference_edas	0.8	55.17	58.06	55.17	39.13	50.00	40.00
Conference_ekaw	0.8	40.00	45.00	41.02	44.89	51.06	46.15
Conference_iasted	0.7	33.33	43.47	33.33	36.84	41.37	34.14
Conference_sigkdd	0.8	58.33	56.00	58.33	40.00	50.00	38.88
confOf_edas	0.8	58.82	64.86	60.60	63.63	66.66	60.86
confOf_ekaw	0.8	66.66	60.60	62.50	59.45	71.79	70.00
confOf_iasted	0.7	62.50	66.66	62.50	42.42	46.15	41.17
confOf_sigkdd	0.7	66.66	66.66	61.53	38.09	47.05	38.09
edas_ekaw	0.8	48.48	50.00	52.94	50.00	65.11	48.00
edas_iasted	0.7	46.15	51.61	50.00	38.59	43.47	32.70
edas_sigkdd	0.8	60.86	60.86	60.86	60.00	51.85	56.25
ekaw_iasted	0.7	52.63	60.86	50.00	31.81	42.42	30.40
ekaw_sigkdd	0.8	66.66	66.66	66.66	63.63	60.00	58.33
iasted_sigkdd	0.8	75.86	68.96	74.07	60.86	75.67	54.16

one-to-one matching form using the linear optimization [4]. It produces the maximal matching that maximizes overall confidence value of the one-to-one matching form alignment ($Align_{lo}$). We have chosen two ontology matching algorithms for each category as they were computing different confidence values for the same concept pair (in some cases, the difference is as high as 0.2). For $\rho = 0.1$, we get two different threshold values 0.7 and 0.8 for different ontology pairs as shown in the table 1. We have excluded both ontology matching algorithms for the category for which we want to generate the reference set.

Table 1 shows the F-measure values for two different implementations of SF as discussed above. From the table 1, it is clear that our framework generates good quality reference set (maximum F-measure being around 88%). From the F-measure values, we can conclude that not only SF selection strategy influences the quality of reference set, but the enabled algorithms (and therefore, their properties) play an important role too. This behavior is consistent and can be observed for multiple ontology pairs of the conference dataset. Obtained results point to an important direction for generating unbiased reference set: the right mix of ontology matching algorithms exploiting different properties with right selection strategy.

Discussion and Future work: In its current avatar, the proposed framework does not model Inter-Algorithm disagreement between ontology matching algorithms exploiting the similar or different properties. The modeling of Inter-Algorithm disagreement may further improve the quality of the generated reference set and reduces the bias in it. The framework does not account for the impact of approach that generates one-to-one matching form on the reference set. Both research questions require further investigation.

The notion of bias, accounted by the proposed framework, is based on a property exploited by a given ontology matching algorithm. Therefore, that property is applicable for all mappings of a reference set. The evaluation exercise of Harrow et al. considers the bias based on the similarity between two participating ontology matching systems and it is mapping specific [2]. If two variants of the same participating system votes for a mapping, it is counted only once.

To generate the output that can be used in real world applications, domain experts need to further validate the generated reference set. Our framework will reduce the efforts required by domain experts in generating silver standard or gold standard. More experiments are needed to further validate the framework with respect to i) the diversity of ontology matching algorithms (e.g., hybrid ontology matching approaches combining and exploiting different properties) and ii) real world ontologies.

References

1. Euzenat, J., Roşoiu, M.E., Trojahn, C.: Ontology matching benchmarks: generation, stability, and discriminability. Journal of web semantics **21**, 30–48 (2013), https://doi.org/10.1016/j.websem.2013.05.002
2. Harrow, I., Jiménez-Ruiz, E., Splendiani, A., Romacker, M., Woollard, P., Markel, S., Alam-Faruque, Y., Koch, M., Malone, J., Waaler, A.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. Journal of biomedical semantics **8**(1), 55 (2017), https://doi.org/10.1186/s13326-017-0162-9
3. Hees, J., Bauer, R., Folz, J., Borth, D., Dengel, A.: Edinburgh associative thesaurus as rdf and dbpedia mapping. In: European Semantic Web Conference. pp. 17–20. Springer (2016), https://doi.org/10.1007/978-3-319-47602-5_4
4. Matousek, J., Gärtner, B.: Understanding and using linear programming. Springer Science & Business Media (2007)
5. Paulheim, H.: Dbpedianyd-a silver standard benchmark dataset for semantic relatedness in dbpedia. In: NLP-DBPEDIA@ ISWC. Citeseer (2013)