Combining Syntactic and Semantic Evidence for Improving Matching over Linked Data Sources

Klitos Christodoulou(✉), Alvaro A.A. Fernandes, and Norman W. Paton

School of Computer Science, University of Manchester,
Oxford Road, Manchester M13 9PL, UK
{christodoulou,alvaro,norm}@cs.man.ac.uk

Abstract. In the context of Linked Data (LD) sources, the ability to traverse links and retrieve further information can be exploited to harvest semantic annotations. Such annotations can, in turn, underpin the inference of semantic correspondences between sources. This paper shows that using semantic annotations as additional evidence of equivalence between schematic representations of LD sources can improve upon the prevalent, purely syntactic approaches. The paper both describes the construction of probabilistic models that yield degrees of belief on the equivalence of the real-world concepts represented by the data and shows how these models are crucial in underpinning a Bayesian approach to assimilating both syntactic evidence (in the form of similarity scores derived by string-based matchers) and semantic evidence (in the form of semantic annotations stemming from LD vocabularies) of equivalence. The paper presents an empirical evaluation of the techniques described. The main finding is confirmation that, with respect to equivalence judgements made by human experts, the use of the contributed techniques incurs significantly fewer discrepancies than purely syntactic approaches.

Keywords: Probabilistic matching · Bayesian updating · Linked data

1 Introduction

The Web of Data (WoD) encourages publishers to make their datasets publicly available. This can lead to a great diversity of publication processes, and inevitably means that resources from the same domain may be described in different ways, using different terminologies. Such heterogeneous representations mean that it can be difficult to identify relationships between published resources, where an understanding of such relationships is useful both for providing an integrated representation of the available data and for linking. Several approaches, from *schema matching* [12] to *ontology alignment* [15], have been

K. Christodoulou—First author has been supported by funding from the UK Engineering and Physical Sciences Research council, whose support we are pleased to acknowledge.

J. Wang et al. (Eds.): WISE 2015, Part I, LNCS 9418, pp. 200–215, 2015.
DOI: 10.1007/978-3-319-26190-4_14

proposed for identifying such candidate relationships (e.g., equivalence). Such techniques typically build on an aggregate of measures of syntactic relationships (such as edit-distance or n-gram intersection) that can be used to hypothesise equivalence. This dependency on syntactic relationships means that decisions tend to suffer from uncertainty. In LD, the fact that resources are described using shared ontologies presents an opportunity to bring together evidence at both the syntactic and semantic levels, i.e., not just names but also semantic annotations that characterise entities at the conceptual level. This paper describes a Bayesian technique for combining syntactic evidence, available in the form of similarity scores computed by string-based matchers, and semantic evidence, available in the form of semantic annotations such as subclass of and equivalent relations that can be formed in, or inferred from, LD ontologies.

Motivating Example. As an example, assume the existence of a LD dataset that describes instance data about music producers, such as, solo artists and groups like "The Beatles". Further, assume that an RDF resource exists for "The Beatles" stating that it is a member of the class *mo:MusicGroup*; *rdf:type(ns1: beatles, mo:MusicGroup)* in the Music Ontology[1]. At the same time, some other music provider models the same instance information about "The Beatles" (potentially under a different URI) stating that a resource for that entity is a member of the class *foaf:Group*; *rdf:type(ns2:beatles, foaf:Group)* using the FOAF vocabulary[2]. Such a scenario is plausible on the WoD since the two resources have been created by different, independent publishers. A system that is interested in merging the two LD datasets needs to deal with such heterogeneity in terminologies by discovering semantic correspondences between the two datasets.

Typically, schema matching techniques utilise knowledge from a formal structure, such as an *ontology description* or a *database schema*, for deriving correspondences of equivalence [2,15]. To make a decision as to the semantic equivalence of the concepts *MusicGroup* and *Group*, assume an approach that applies a set of string-based matchers (such as edit-distance and n-gram) over the localnames of *mo:MusicGroup* and *foaf:Group*. Such algorithms typically use a similarity score from the interval [0,1] as a confidence measure for discovered correspondences. Note that, in addition to their syntactic relationship, a semantic relation exists, stating that *mo:MusicGroup* is subsumed by *foaf:Group*. We suggest that such relations can be used as additional knowledge to improve the decision making of matching techniques beyond the use of syntactic matchers alone.

Summary of Contributions. This paper describes a probabilistic approach for combining evidence from syntactic matchers with semantic annotations modelled as *degrees of belief* on the existence of semantic correspondences of equivalence. The following questions motivated this study. Which semantic annotations can be usefully be taken as additional evidence for the purposes of postulating construct equivalence? How can we reason about syntactic and semantic evidence using probabilistic models? How can we incrementally assimilate different

[1] http://purl.org/ontology/mo/.
[2] http://xmlns.com/foaf/0.1/.

kinds of evidence? In seeking solutions to these questions, this paper contributes the following: (a) a methodology that uses kernel density estimation for deriving likelihoods from similarity scores computed by string-based matchers; (b) a methodology for deriving likelihoods from semantic relations (e.g., rdfs: subClassOf, owl:equivalentClass) that are retrieved by dereferencing URIs in LD ontologies; (c) a methodology for aggregating evidence of conceptual construct equivalence from both string-based matchers and semantic annotations; and (d) an empirical evaluation of our approach grounded on the judgements of experts in response to the same kinds of evidence.

The remainder of the paper is structured as follows. Section 2 presents an overview of the developed solution. Section 3 describes the methodology used for deriving probability distributions over similarity scores from string-based matchers, along with a methodology for deriving likelihoods from semantic knowledge defined in LD ontologies. Bayesian updating, as a technique for the incremental assimilation of evidence, is introduced in Sect. 4. Section 5 presents an empirical evaluation of the methodology complemented by a discussion of results. Section 6 reviews related work, and Sect. 7 concludes.

2 Overview of Solution

Given a conceptual description of a *source* and a *target* LD dataset, denoted by S and T, respectively, a *semantic correspondence* of equivalence is a triple $\langle c_S, c_T, P(c_S \equiv c_T|E)\rangle$, where $c_S \in S$ and $c_T \in T$ are constructs (i.e., Classes) from the datasets, and $P(c_S \equiv c_T|E)$ is the conditional probability representing the degree of belief (from now on referred to as dob) in the equivalence ($\equiv$) of the constructs given the pieces of evidence $(e_1, ..., e_n) \in E$. Section 4 describes in detail how to compute the conditional probability using Bayes' theorem. Our approach distinguishes two types of knowledge: (a) *syntactic knowledge*, in the form of strings that are local-names of resources' URIs; and (b) *semantic knowledge*, such as structural relations between entities, either internal to a vocabulary or across different LD vocabularies, e.g., relations such as subclass of and equivalence. Table 1 summarises the types of knowledge construed by our approach as sources of evidence. The set TE is the set of all semantic annotations we consider as evidence, where the subsets EE and NE comprise the assertions that can be construed as *direct* evidence of equivalence and non-equivalence, respectively.

To collect syntactic evidence (represented by the set LE), given two sources, our approach extracts local-names from the URIs of every pair of constructs $\langle c_s, c_t\rangle$ and then derives their pair-wise string-based similarity. Two string-based metrics are used, viz., *edit-distance* (denoted by ed) and *n-gram* (denoted by ng) [15]. Section 3.1 elaborates on how probability distributions can be constructed for each matcher. To collect semantic evidence, our approach dereferences URIs to get access to annotations from the vocabularies that define the resource. For example, the subsumption relation $c_S \sqsubseteq c_T$ is taken as semantic evidence. Section 3.2 elaborates on an approach to constructing probability distributions for each kind of semantic evidence in RDFS/OWL vocabularies.

Table 1. Syntactic and semantic evidence utilised by the technique.

Type		ID	Description	Evidence rule
Syntactic evidence (LE)	-	SLN	similar-local-name	$string\ similarity(c_T, c_S)$
Semantic evidence (TE)		SU	same-URI	$string\ equality(URI_S, URI_T)$
	-	SB	subsumed-by	$c_S \sqsubseteq c_T$
	EE	SA	same-as	owl:sameAs(c_S, c_T)
		EC	equivalent-class	owl:equivalentClass(c_S, c_T)
		EM	exact-match	skos:exactMatch(c_S, c_T)
	NE	DF	different-from	owl:differentFrom(c_S, c_T)
		DW	disjoint-with	owl:disjointWith(c_S, c_T)

3 Constructing Likelihoods for Evidence

To assimilate different kinds of evidence, some bootstrapping is needed that will allow the computation of the likelihoods necessary for the calculation of a dob on construct equivalence, as captured by the posterior $P(c_S \equiv c_T|E)$ given both syntactic and semantic evidence. This section describes a principled methodology for constructing probability distributions from similarity scores returned by string-based matchers, as well as a procedure for deriving likelihoods for each type of semantic evidence in Table 1.

3.1 Similarity Scores to Degrees of Belief

We call *syntactic evidence* the likelihoods derived from *similarity scores* produced by string-based matchers. We study the behaviour of each matcher (in our case ed and ng) to derive these likelihoods as follows:

1. From the datasets made available by the Ontology Alignment Evaluation Initiative (OAEI)[3], we observed the available ground truth on whether a pair of local-names, denoted by (n, n'), aligns.
2. We assume the existence of a continuous random variable, X, in the bounded domain [0,1], for the similarity scores returned by each matcher μ, where $\mu \in$ ed, ng. Our objective is to model the behaviour of each matcher in terms of a probability density function (PDF) $f(x)$ over the similarity scores it returns (we refer to them as observations).
3. To empirically approximate $f(x)$ for each matcher we proceed as follows:
   (a) We ran each matcher μ independently over the set of all local-name pairs (n, n') obtained from (1).
   (b) For each pair of local-names, we observed the independent similarity scores returned by the matcher when (n, n') agrees with the ground truth. These are the set of observations $(x_1, ..., x_i)$ from which we estimate $f(x)$ for the equivalent case.

[3] http://oaei.ontologymatching.org.

4. The observations $x_1, ..., x_i$ obtained are used as inputs to the non-parametric technique known as kernel density estimation (KDE) (using a Gaussian kernel[4]) [3] whose output is an approximation $\hat{f}(x)$ for both ed and ng for both the equivalent and non-equivalent cases.

We interpret the outcome of applying such a PDF to syntactic evidence as the likelihood of that evidence. More formally, and as an example, $PDF_{\underset{ed}{\equiv}}(\mathsf{ed}(n, n'))$ $= P(\mathsf{ed}(n, n')|c_S \equiv c_T)$, i.e., given a pair of local-names (n, n') the PDF for the ed matcher in the equivalent case $PDF_{\underset{ed}{\equiv}}$ yields the likelihood that the similarity score $\mathsf{ed}(n, n')$ expresses the equivalence of the pair of concepts (c_S, c_T) that (n, n'), resp., denote. Correspondingly, for the non-equivalent case, and for ng in both the equivalent and non-equivalent cases (Fig. 1).

The probability distributions derived by this process are shown in Fig. 2(a) and (b) for ed and in Fig. 2(c) and (d) for ng. The procedure described can be used to study the behaviour of any matcher that returns similarity scores in the interval $[0, 1]$. Note that the PDFs obtained by the method above are *derive-once*, *apply-many* constructs. Assuming that the sample set used for training remains representative, and given that the behaviour of matchers ed and ng is fixed and deterministic, the PDFs need not be recomputed.

3.2 Semantic Evidence to Degrees of Belief

We call *semantic evidence* the likelihoods derived from *semantic annotations* obtained from the WoD. We first retrieved the semantic annotations summarised in Table 1. The set TE is the set of all such evidence, TE = {SU, SB, SA, EC, EM, DF, DW}. We formed the subsets $EE \subset TE$ = {SA, EC, EM} and $NE \subset TE$ = {DF, DW} comprising assertions that can be construed as *direct* evidence of equivalence and non-equivalence, respectively.

To derive probability distributions for semantic evidence, we proceeded as follows:

1. We assume the existence of a Boolean random variable, for each type of semantic evidence in Table 1, with domain $\{true, false\}$.
2. Using the vocabularies available in the Linked Open Vocabularies (LOV)[5] collection.
   (a) We collected and counted pairs of classes and properties that share direct or indirect assertions of equivalence or non-equivalence for all the assertions in TE and NE using SPARQL queries. For example:

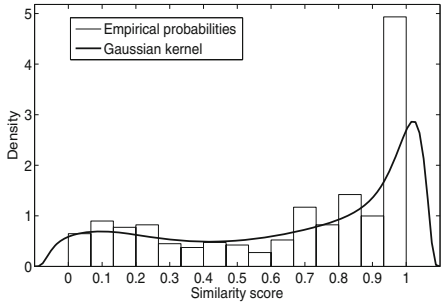
[4] Gaussian kernel was used due to its mathematical convenience. Note that any kernel other than Gaussian can be applied, however, the shape of the distribution may differ depending on the kernel characteristics.

[5] http://lov.okfn.org/dataset/lov/.

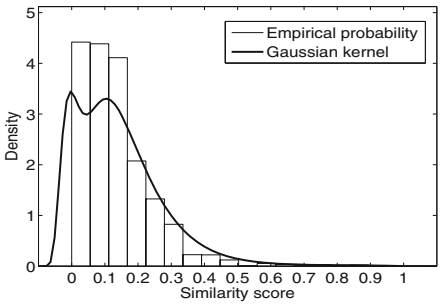
```
SELECT DISTINCT ?elem1 ?elem2
WHERE {
    {?elem1 a rdfs:Class .} UNION {?elem1 a owl:Class .}
    ?elem1 ?p ?elem2 .
    FILTER (?p = owl:equivalentClass && !isBlank(?elem2))}
```

(b) From the set of pairs derived by the assertions in TE and NE, we counted assertions that can be construed as *evidence* of equivalence or non-equivalence for each pair, grouping such counts by kind of assertion (e.g., subsumed-by (SB), etc.)

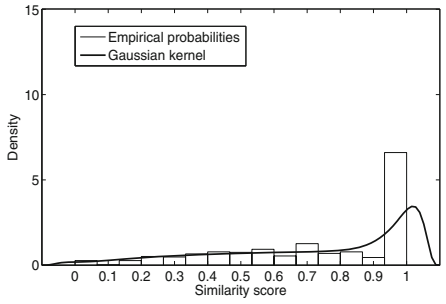
3. We used the sets of counts obtained in the previous step to build contingency tables (e.g., see Table 2) from which we can directly derive the probability mass functions (PMFs) for each kind of semantic evidence for both the equivalence and non-equivalent cases.



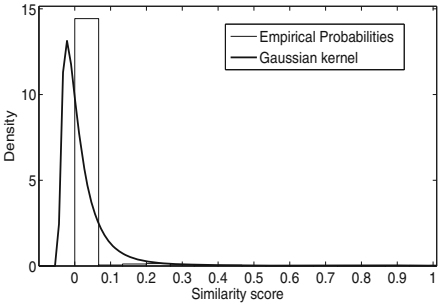
(a) Edit-distance matcher behaviour (equivalent case).



(b) Edit-distance matcher behaviour (non-equivalent case).



(c) N-gram matcher behaviour (equivalent case).



(d) N-gram matcher behaviour (non-equivalent case).

Fig. 1. Illustration of probability distributions for each matcher over $[0, 1]$.

The PMFs obtained through the steps above are also *derive-once*, *apply-many* constructs, but since the vocabulary collection from which we draw our sample is dynamic, we might wish to view them as *derive-seldom*, *apply-often*.

Table 2. Example of a contingency table. The likelihood $P(\mathbf{EC}(n, n')|c_S \equiv c_T)$ is estimated by the fraction 305/396.

Contingency table	Semantic evidence		
	EC	¬EC	Total
$c_S \equiv c_T$	305	91	396
$c_S \not\equiv c_T$	0	2552	2552
Total	305	2643	2948

We interpret the outcome of applying such a PMF to semantic evidence as the likelihood of that evidence. More formally, and as an example, $PMF_{\underset{\mathsf{EC}}{\equiv}}(\mathsf{EC}(u, u')) = P(\mathsf{EC}(u, u')|c_S \equiv c_T)$, i.e., given the existence of an assertion that a pair of URIs (u, u') have an equivalence relation, the probability mass function for this kind of assertion in the equivalent case $PMF_{\underset{\mathsf{EC}}{\equiv}}$ yields the likelihood that the assertion $\mathsf{EC}(u, u')$ expresses the equivalence on the pair of constructs (c_S, c_T) that (u, u'), resp., denote. Correspondingly, for the non-equivalence case and for all other kinds of semantic evidence (e.g., SB, etc.) in both the equivalent and non-equivalent cases.

4 Assimilating Evidence Using Bayesian Updating

The purpose of deriving likelihood models as described in Sect. 3 is to enable the evidence to be combined in a systematic way using Bayesian updating. The procedure for doing so is now described, where the benefits of the procedure are discussed in Sect. 5.

We denote with S and T, resp., the structural summaries (an ontology or a structural summary derived by an approach like [4]) that describe the structure of a *source* and a *target* LD source over which we wish to discover semantic correspondences. Given a pair of constructs $c_S \in S$ and $c_T \in T$ our objective is to derive a dob on the postulated equivalence of a pair of constructs (denoted by H), given pieces of evidence $e_1, ..., e_n \in E$. To reason over our hypothesis, we model it as a conditional probability $P(H|E)$ and apply Bayes' theorem to make judgements on the equivalence of two constructs. In its simplest form, Bayes' theorem states that[6],

$$P(H|E) = \frac{P(E|H)\ P(H)}{P(E)}. \tag{1}$$

Our hypothesis can take one of two states: $P(H) = \{P(c_S \equiv c_T), P(c_S \not\equiv c_T)\}$, i.e., it is a *Boolean hypothesis*. The prior probability, e.g., $P(H) = P(c_S \equiv$

[6] Informally, the theorem states that the hypothesis given the evidence (so called posterior) is equal to the ratio between the product of the dob in the evidence given the hypothesis (what we called likelihood in Sect. 3) and the dob in the hypothesis (so called prior) divided by the dob in the evidence.

c_T), is the dob in the absence of any other piece of evidence (we assume a uniform distribution). Thus, for the two possible outcomes our hypothesis can take, $N = 2$, the prior probability that one of the outcomes is observed is given by $1/N$. The probability of the evidence, $P(E)$, can be expressed using the law of total probability [9], i.e., $P(E) = P(E|c_S \equiv c_T)\ P(c_S \equiv c_T) + P(E|c_S \not\equiv c_T)\ P(c_S \not\equiv c_T)$. To use Bayes' theorem for deriving a dob on the hypothesis given the available evidence, it is essential to estimate the likelihoods for each evidence: (i.e., $P(E|c_S \equiv c_T)$, and, $P(E|c_S \not\equiv c_T)$). For semantic evidence, the likelihoods are estimated from the contingency tables constructed in Sect. 3.2. For continuous values, like similarity scores, the constructed PDFs for each matcher from Sect. 3.1 are used to estimate the conditional probabilities for the likelihoods. To determine these likelihoods, we integrate the PDF over a finite region $[a, b]$, namely $P(a \leq X \leq b) = \int_a^b f(x)\,dx$, where the density $f(x)$ is computed using KDE with a *Gaussian* kernel.

The idea behind *Bayesian updating* [16], is that once the posterior e.g., $P(c_S \equiv c_T|E)$ is computed for some evidence, $e_1 \in E$, a new piece of evidence $e_2 \in E$, leads us to compute the impact of e_2 by taking the previously computed posterior as the new prior. Given the ability to compute likelihoods for both syntactic and semantic evidence, we can use Bayesian updating to compute a dob on the equivalence of (pairs of constructs in) two structural summaries S and T. To demonstrate this with a concrete example, let $P^{(e_1,\ldots,e'_n)}$ denote the dob that results from having assimilated the evidence sequence $(e_1, \ldots, e_n)$. The initial prior is therefore denoted by $P^{()}$, and if $(e_1, \ldots, e_n)$ is the complete evidence sequence available, then $P^{(e_1,\ldots,e'_n)}$ is the final posterior. We proceed as follows:

i. We set the initial prior according to the principle of indifference between the hypothesis that $P(c_S \equiv c_T)$ and its negation, so $P^{()} = 0.5$.
ii. We collect the local-name pairs from the structural summaries S and T.
iii. We run `ed` on the local-name pairs and, using the probability distributions derived using the methodology described above (Sect. 3.1), compute the likelihoods for each pair and use Bayes' rule to calculate the initial posterior $P^{(ed)}$.
iv. We run `ng` on the local-name pairs and, using the probability distributions derived using the methodology described above (Sect. 3.1), compute the likelihoods for each pair and use Bayes' rule to calculate the next posterior $P^{(ed,ng)}$. Note that this is the dob given the syntactic evidence alone, which we denote more generally by $P^{(syn)}$.
v. To get access to semantic annotations that span a variety of LD ontologies, we dereference every URI in S and T to collect the available semantic annotations e.g., $SB(c_S \subseteq c_T)$.
vi. Using the methodology described above (Sect. 3.2), we compute, one at a time, the likelihoods for the available semantic evidence, each time using Bayes' rule to calculate the next posterior (e.g., $P^{(ed,ng,SB,\ldots)}$), so that once all the available semantic evidence is assimilated, the final posterior, which we denote more generally by $P^{(syn,sem)}$, is the dob on $c_S \equiv c_T$, where, $c_S \in S \wedge c_T \in T$.

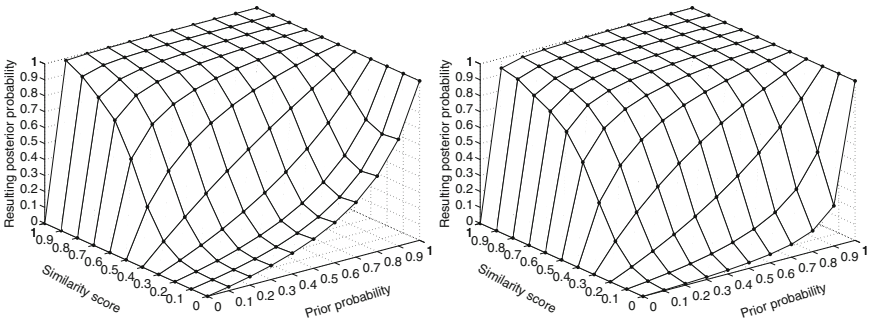
Before carrying out the empirical evaluation of this approach using syntactic and semantic evidence described in Sect. 5, we studied analytically, using Bayes's theorem, the effect of each piece of evidence independently. Given a series of initial prior probabilities in the range of $[0, 1]$ and the evidence likelihoods (see Sect. 3) we computed the posterior probabilities given each piece of evidence. Figure 2(a) and (b) show how the posteriors $P(c_s \equiv c_t|ed(c_s, c_t) = s)$, and, $P(c_s \equiv c_t|ng(c_s, c_t) = s)$, resp., are updated when the available evidence is similarity scores computed by the string-based matchers ed and ng. As an example, consider Fig. 2(a), and assume that we are given a prior probability of $x = 0.5$ and a similarity score that is $y < 0.5$, ed will cause the updated posterior probability to fall relatively more. In this case, if the similarity score is $y = 0.2$, the posterior probability drops to $z = 0.2$. In the case of ng, using identical values as previously, the posterior probability drops to $z = 0.36$, which means that ng causes a small decrease in the posterior. In a similar fashion, the independent behaviours of different kinds of semantic evidence have been studied. For example, Fig. 2(c) shows how the posterior is updated when there is direct evidence that a pair of classes stand in a subsumption relationship (i.e., SB). A subsumption relation may indicate that the constructs are more likely to be related than to be disjoint and a low prior is therefore increased. Similarly, Fig. 2(d) shows how the posterior is affected when a pair of constructs stand in an equivalence relation (i.e., EC). This is considered enough evidence to significantly increase a low prior to close to 1; meaning that constructs are more probably equivalent than if that evidence had not been available.

Having observed how different posterior probabilities are updated in the presence of individual pieces of evidence, in Sect. 5 we empirically assess whether the incorporation of semantic evidence from LD ontologies can improve on construct equivalence judgements obtained through syntactic matching alone.

5 Experimental Evaluation

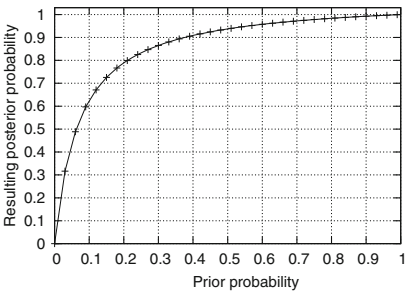
The evaluation of our approach was based on the idea of emulating the construct equivalence judgements produced by human experts in the presence of different kinds of syntactic and semantic evidence[7]. The judgements derived from experts are then compared with the judgements derived by the Bayesian updating approach discussed in Sects. 3 and 4. This section describes an experimental scenario that has a twofold purpose: (a) to compare how well the Bayesian assimilation of syntactic evidence alone performs against the aggregation of syntactic evidence followed by a predefined function, specifically average (AVG) which is commonly used in existing matching systems [2,15], and (b) to observe empirically whether the incorporation of semantic evidence can improve on construct equivalence judgements obtained through syntactic matching alone.

[7] The survey was distributed and completed by 15 human participants all experts in solving data integration tasks, such as schema matching and mapping.

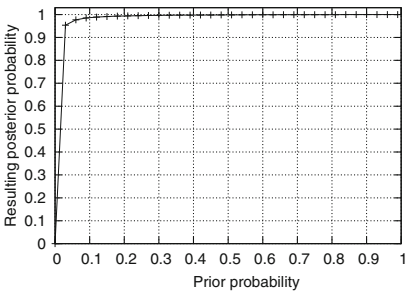


(a) Edit-distance matcher.

(b) N-gram matcher.



(c) Evidence for subsumption.



(d) Evidence of equivalence.

Fig. 2. Effect on the posterior probabilities using particular evidence on different prior probabilities.

5.1 Experimental Setup

To evaluate the application of Bayes's theorem for assimilating different kinds of evidence, the experimental evaluation was grounded on the rational decisions made by *human experts* on data integration and ontology alignment when judging whether a pair of constructs is postulated to be equivalent given both syntactic and semantic evidence as construed in this paper. For the purposes of the experiment, a set of pairs of constructs from different LD ontologies was collected, making sure that different combinations of syntactic and semantic evidence (as in Table 1) were present or absent. To obtain testimonies from the human experts, a survey was designed based on the collected set of pairs of constructs, asking the experts to make judgements on the equivalence of such pairs. Testimonies have been recorded on a discretisation scale [5], as follows: {Definitely equivalent} mapped to a dob of 1.0; {Tending towards being equivalent} mapped to a dob of 0.75; {Do not know} mapped to a dob of 0.5; {Tending towards being not-equivalent} mapped to a dob of 0.25; and {Definitely not-equivalent} mapped to a dob of 0. By observing different pairs of constructs from real ontologies, approximately 40 common combinations of syntactic and semantic evidence have been identified. For each combination, a question was designed to obtain individual

testimonies from each responder. Individual testimonies from each question were aggregated using a weighted average, based on the confidence assigned to each item [5]. The aggregated degrees of belief obtained from the survey are treated as an approximation of the experts' confidence on construct equivalence given certain pieces of syntactic and semantic evidence and act as a gold standard.

Datasets. For the purposes of the experiment, the Bayesian technique was evaluated over the class hierarchies of ontologies made available by the OAEI - Conference track, which have been designed independently by different processes but all belonging to the domain of conference organisation. Note also that these ontologies share no semantic relations between them. Since our technique assumes such relations for use as semantic evidence, we made some of these cross-ontology semantic relations explicit using BLOOMS[8]; a system for discovering rdfs:subClassOf and owl:equivalentClass relations between LD ontologies [7]. We note that the contributions reported in this paper are independent of BLOOMS, in that they can be used regardless of the sources of semantic annotations. We found that the LOD cloud at the conceptual level still lacks the abundance of cross-ontology links that, most agree, will one day characterise the SW. We have therefore used BLOOMS to induce some more cross-ontology links in a principled manner. The results reported in this paper consider a single pair of ontologies from the conference track, viz., `ekaw` (denoted by S) and `conference` (denoted by T).

Expectation Matrix. Given a pair of classes from the class hierarchies of the input ontologies and given the available kinds of evidence, both syntactic and semantic, a dob was assigned for each pair on the basis of the experts' testimonies. More formally, we constructed a $n \times m$ structure referred to from now on as the *expectation matrix* and denoted as M_{exp}, where $n = |S|$ and $m = |T|$. The element e_{jk} in the jth row and the kth column of M_{exp} denotes the dob derived from the expert survey between the jth construct in S and the kth construct in T according to the pieces of evidence present or absent.

Evaluation Metric. Let $p_1, p_2, ..., p_n$ be the degrees of belief derived for each pair of classes from the ontologies by either the average aggregated scheme or the Bayesian assimilation, and $a_1, a_2, ..., a_n$ be the corresponding degrees of belief in the expectation matrix just described. We compute the mean-absolute error, $MAE = (|p_1 - a_1| + ... + |p_i - a_i|) \div n$ where $|p_1 - a_1|$ is an *individual error* of a pair and n is the total number of such errors.

5.2 Evaluation Methodology

Traditional matching approaches (e.g., COMA [1]) exploit different pieces of evidence, mostly from string-based matchers, to assess the similarity between constructs in ontologies or in database schemas. Such approaches combine similarity scores computed independently, typically using averages. For this evaluation the antagonist to our Bayesian approach is considered a process that independently runs matchers `ng` and `ed` on the local-names of classes from ontologies S and T,

[8] BLOOMS was configured with a high threshold, viz., > 0.8..

and produces an average of the similarity scores. The aggregated result of this computation is a matrix M_{avg}. The next step is to measure how close the derived predictions are to the degrees of belief obtained by the experts' testimonies. In doing so, we used MAE as the performance measure since it does not exaggerate the effect of outliers [6]. The result from computing the error between M_{avg} and the expectation matrix M_{exp} is denoted by δ_{avg}.

Similarly, the Bayesian assimilation technique (as described in Sect. 4) was used (instead of an average) to assimilate the evidence computed by the string-based matchers on pairs of local-names. The result of this computation is a matrix M_{syn}, where $n = |S|$ and $m = |T|$. The element e_{jk} in the jth row and the kth column of M_{syn} denotes the posterior probability $P^{(syn)}$ between the jth class in S and the kth class in T according to the syntactic evidence derived from the string-based matchers ed and ng. The next step is to measure how close the predictions from M_{syn} are to the expectation matrix M_{exp}. The result is denoted by δ_{syn}.

To assess whether semantic evidence can improve on construct equivalence judgements that use averaging alone to aggregate syntactic evidence, we first used BLOOMS [7] to make explicit the cross-ontology semantic relations and used this as semantic evidence. In the light of this new evidence, the Bayesian assimilation technique updates the posterior probabilities $P^{(syn)}$ for each pair of classes in M_{syn} accordingly. The result of this process is a new matrix $M_{syn,sem}$ with the same dimensions as M_{syn}, where, the posterior probabilities for the elements e_{jk} reflect both syntactic and semantic evidence, $P^{(syn,sem)}$. Again we denote by $\delta_{syn,sem}$ the error calculated between $M_{syn,sem}$ and the expectation matrix M_{exp}. Finally, to complete the evaluation, the individual absolute errors used for the calculation of δ_{avg}, δ_{syn}, and $\delta_{syn,sem}$ have been examined. The results of the evaluation are now discussed.

5.3 Results and Discussion

Exp. 1: AVG scheme vs. Bayesian Syntactic. The MAE error computed for the average aggregation scheme against the expectation matrix was $\delta_{avg} = 0.1079$ whereas the error as a result of assimilating syntactic evidence using the Bayesian technique was $\delta_{syn} = 0.0698$. The difference of 0.0381 between the two errors can be expressed in percentage terms as 35.32 %. To further understand the difference in errors, we measured the individual *absolute* errors that fall into each of four regions of interest as these are shown in Fig. 3(a). They correspond to the following minimum bounding rectangles, resp., Region 1 lies below the $y = x$ error line where AVG error >> Bayesian error and is the rectangle defined by $y = 0.2$; Region 2 lies above the $y = x$ error line where AVG error << Bayesian error and is the rectangle defined by $x = 0.2$; Region 3 lies below the $y = x$ error line where AVG error > Bayesian error and is the rectangle defined by $y > 0.2$; and Region 4 lies above the $y = x$ error line where AVG error < Bayesian error and is the rectangle defined by $x > 0.2$. We note that the larger the cardinality of Region 1, the more significant is the impact of using semantic annotations as we propose.

Table 3. AVG scheme vs. Bayesian syntactic.

No.	Region	Count	Perc. (%)
1	$R_{avg>>B_{syn}}$	3833	87.49
2	$R_{avg<<B_{syn}}$	215	4.90
3	$R_{avg>B_{syn}}$	31	0.70
4	$R_{avg<B_{syn}}$	302	6.89

Table 4. AVG scheme vs. Bayesian syntactic & semantic.

No.	Region	Count	Perc. (%)
1	$R_{avg>>B_{syn,sem}}$	125	71.43
2	$R_{avg<<B_{syn,sem}}$	43	24.57
3	$R_{avg>B_{syn,sem}}$	2	1.14
4	$R_{avg<B_{syn,sem}}$	5	2.85

Table 5. Bayesian syntactic vs. Bayesian syntactic & semantic.

No.	Region	Count	Perc. (%)
1	$R_{B_{syn}>>B_{syn,sem}}$	124	89.21
2	$R_{B_{syn}<<B_{syn,sem}}$	9	6.48
3	$R_{B_{syn}>B_{syn,sem}}$	5	3.60
4	$R_{B_{syn}<B_{syn,sem}}$	1	0.72

For the traditional aggregation scheme that produced M_{avg} we counted 3833 matches with individual errors greater than the analogous individual errors derived by the Bayesian technique that produced M_{syn}. The use of Bayesian aggregation significantly outperformed (i.e., has smaller individual errors than) the use of AVG aggregation scheme for 87.49 % of the total. Table 3 summarises the results for each region showing how many individual errors are located in each of the regions of interest in both absolute terms and relative to the total.

Exp. 2: AVG scheme vs. Bayesian Syn. & Sem. To evaluate our hypothesis whether semantic annotations can improve outcomes we compared the aggregated errors denoted by δ_{avg} and $\delta_{syn,sem}$. The mean absolute error $\delta_{syn,sem} = 0.1259$ is lower than $\delta_{avg} = 0.1942$ with a difference of 0.0683 or 35.15 %. Figure 3(b) plots the individual errors for pairs of classes that have some semantic relation between them. We are interested on cases where the individual errors for the Bayesian technique are smaller than the AVG scheme. In particular, the points that lie mostly between 0.1 and 0.3 on the x-axis and below the $y = x$ error line. For 71.43 % of the total matches that have some semantic evidence the Bayesian technique produces results closer to the testimonies, with individual errors that mostly lie in that region. Table 4 summarises the results for each region showing how many individual errors are located in each of the regions of interest in both absolute terms and relative to the total.

Exp. 3: Bayesian Syn. vs. Bayesian Syn. & Sem. Similarly to Exp.2, we compared the aggregated errors denoted by δ_{syn} and $\delta_{syn,sem}$ considering only individual errors that have some semantic evidence. Again in this case $\delta_{syn,sem} = 0.1259$ is closer to the expectation matrix than $\delta_{syn} = 0.2768$ with a difference of 0.1509 or 54.5 %. The results of this experiment are summarised in Table 5.

The points of interest in this experiment are the ones where the individual errors for $B_{syn,sem}$, that considers both syntactic and semantic evidence, are smaller than B_{syn}. For 89.21 % of the total matches discovered, that have some semantic evidence, $B_{syn,sem}$ outperforms the configuration of the Bayesian scheme that utilises syntactic evidence alone, i.e., B_{syn}.

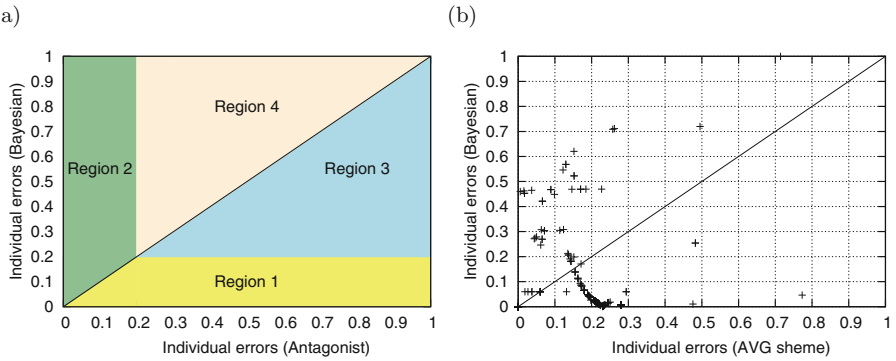


Fig. 3. (a) Shows the regions of interest, (b) Individual errors Bayesian against AVG scheme.

Overall, the experimental evaluation showed that the Bayesian assimilation of syntactic and semantic evidence delivers, in most cases better judgements of construct equivalence than the use of syntactic matchers alone i.e., than the state-of-the-art in matching. The aim of the experiment was to evaluate the Bayesian approach on how well it performed on aggregating different pieces of syntactic and semantic evidence against AVG a common aggregation strategy used in traditional matching approaches. Alignments provided by the OAEI group are tailored towards evaluating approaches that make classification decisions for discovering ontology alignments and are not suitable for judging individual aggregated confidence degrees of belief derived by the approaches. To the best of our knowledge there are no established benchmarks for doing so. Therefore, we consulted human experts for the construction of the baseline used for the evaluation.

6 Related Work

A variety of strategies have been proposed in the literature for solving the problem of combining different pieces of evidence about matches, some examples are: average, weighted average, min, max and sigmoid functions [10]. However, it falls on users to tune or select the appropriate aggregation method manually according to the problem in hand. In contrast, the Bayesian assimilation of

evidence technique can be used as an alternative aggregation strategy for assimilating any piece of evidence, complementing typical aggregation strategies used by state-of-the-art schema and ontology matching systems [2,12,15]. When the appropriate probability distributions are made available, the approach presented in this paper can be used as a generic aggregation strategy that is not tied to any specific domain. Sabou et al. [13] presented an ontology matching paradigm that makes use of additional external background knowledge that is made available from ontologies from the Semantic Web. The proposal in our paper makes use of additional semantic annotations from LD ontologies as evidence with the aim of improving the decision making of different matchers that mostly work on syntax. In another note, approaches for discovering semantic relations from ontologies e.g., [14] can be used to provide input to our Bayesian approaches to further improve the accuracy, thus improving the decision making of matching approaches. The uncertainty in the decisions made by different *matchers* has also been observed in [8], where a similarity matrix that describes the outcome of some matcher is modelled as two probability distributions. An alternative statistical analysis is used to model the similarity scores distribution returned by each matcher that uses the parametric beta-distribution to estimate the underlying probability. The proposal in our paper, however, makes no assumptions about the shape or parameters of the underlying distribution, and uses a non-parametric statistical analysis technique, based on kernel density estimation, to approximate the probability distributions for each matcher using the sampled data.

7 Conclusions

The WoD can be seen as vibrant but challenging: vibrant because there are numerous publishers making valuable data sets aware for public use; challenging because of inconsistent practises and terminologies in a setting that is something of a free-for-all. In this context, it is perhaps easier to be a publisher than a consumer. As a result, there is a need for tools and techniques to support effective analysis, linking and integration in the web of data [11]. The challenging environment means: (i) that there are many different sources of evidence on which to build; (ii) that there is a need to make the most of the available evidence; and (iii) that it is not necessarily easy to do (ii). This paper has described a well-founded approach to combining multiple sources of evidence of relevance to matching, namely syntactic matchers and semantic annotations. The findings from the empirical evaluation suggested that the Bayesian aggregation scheme has let to improved decision making of close to 90 % of the total matches when assimilating just syntactic evidence, and the confidence of close to 70 % of the matches that had some semantic evidence has been improved in the light of available semantic evidence. Overall, the suggested approach can be used as a generic methodology for assimilating different kinds of evidence as they become available, or as a method that complements existing aggregation strategies for matching systems.

References

1. Aumueller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and ontology matching with coma++. In: SIGMOD Conference, pp. 906–908 (2005)
2. Bernstein, P., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. Proc. VLDB Endowment **4**(11), 695–701 (2011)
3. Bowman, A.W., Azzalini, A.: Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations: The Kernel Approach with S-Plus Illustrations. OUP, Oxford (1997)
4. Christodoulou, K., Paton, N.W., Fernandes, A.A.A.: Structure inference for linked data sources using clustering. In: EDBT/ICDT Workshops, pp. 60–67 (2013)
5. de Vaus, D.: Surveys in Social Research. Research methods/Sociology. Taylor & Francis (2002)
6. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. Int. J. Forecast. (IJF) **22**(4), 679–688 (2006)
7. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010)
8. Marie, A., Gal, A.: Managing uncertainty in schema matcher ensembles. In: Prade, H., Subrahmanian, V.S. (eds.) SUM 2007. LNCS (LNAI), vol. 4772, pp. 60–73. Springer, Heidelberg (2007)
9. Papoulis, A.: Probability, Random Variables and Stochastic Processes, 3rd edn. McGraw-Hill Companies, New York (1991)
10. Peukert, E., Maßmann, S., König, K.: Comparing similarity combination methods for schema matching. GI Jahrestagung **1**, 692–701 (2010)
11. Polleres, A., Hogan, A., Harth, A., Decker, S.: Can we ever catch up with the web? Semantic Web **1**(1–2), 45–52 (2010)
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. **10**(4), 334–350 (2001)
13. Sabou, M., d'Aquin, M., Motta, E.: Exploring the semantic web as background knowledge for ontology matching. J. Data Semant. **11**, 156–190 (2008)
14. Sabou, M., d'Aquin, M., Motta, E.: SCARLET: semantiC relAtion discoveRy by harvesting onLinE onTologies. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 854–858. Springer, Heidelberg (2008)
15. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013)
16. Spragins, J.: A note on the iterative application of bayes' rule. IEEE Trans. Inf. Theor. **11**(4), 544–549 (2006)