Knowledge Media Institute

The Open University

Towards Linked Data Fact Validation through Measuring Consensus

Shuangyan Liu, Mathieu d'Aquin, Enrico Motta

Knowledge Media Institute (KMi), The Open University, UK

Shuangyan.Liu@open.ac.uk

Linked Data Quality Workshop @ ESWC

Linked Data Quality Assessment

- Data quality definition
  - "multi-dimensional construct with a popular definition 'fitness for use'" (Zaveri et al. 2012)
- New challenges
  - Openness of the linked data
  - Diversity of the data
  - Dynamic set of autonomous data sources and publishers
- Linked data quality dimensions (Zaveri et al. 2012)
  - Accessibility dimensions
  - Representational dimensions
  - Contextual dimensions
  - Intrinsic dimensions
- **Semantic accuracy** dimension
  - "the degree to which data values correctly represent the real world facts" (Zaveri et al. 2012)

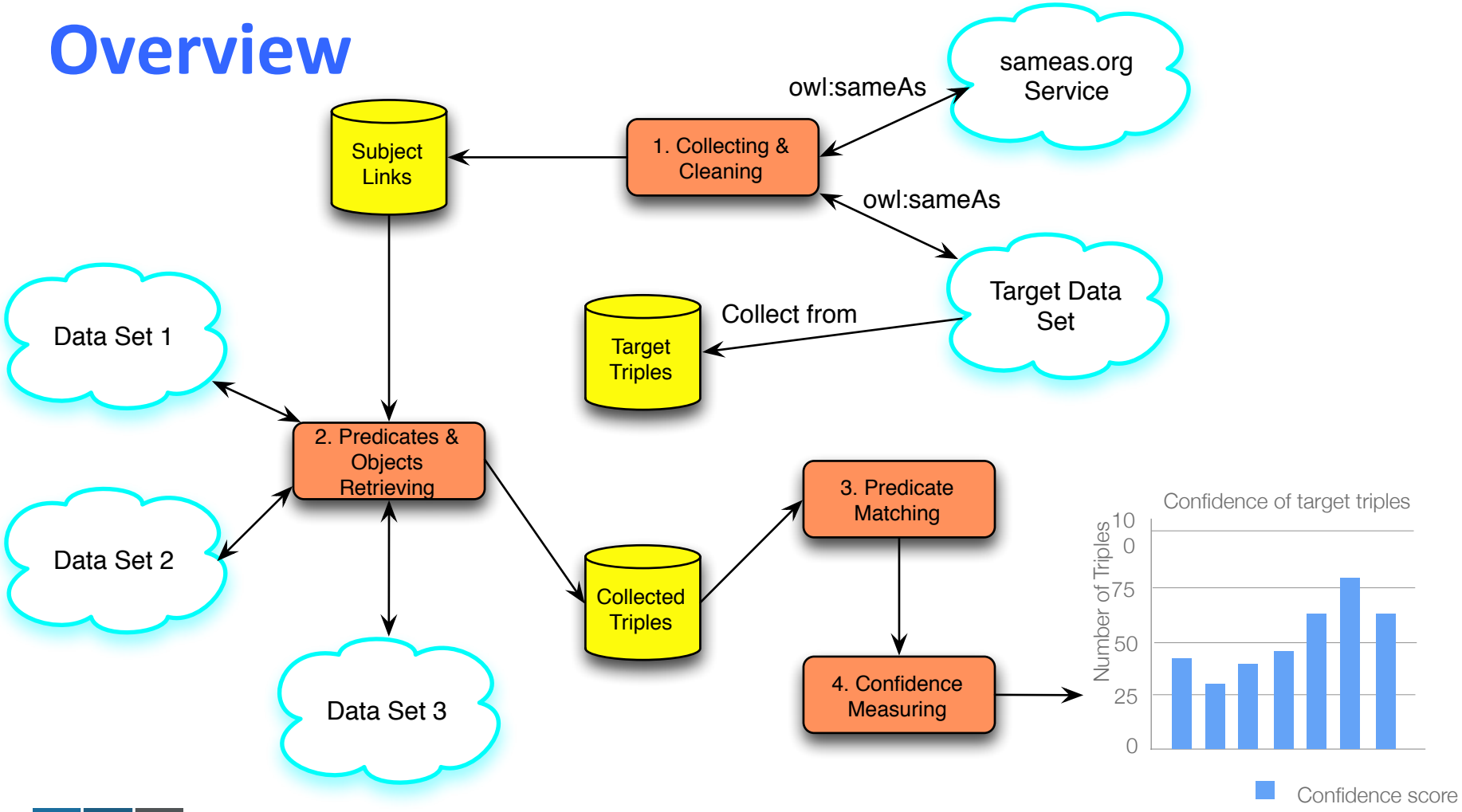
Problem Statement

- Assessing the semantic accuracy of the linked data
  - How to identify inaccurate facts or values in a linked dataset
- Existing approaches for detecting inaccurate values
  - **SWIQA** (Furber and Hepp, 2011)
    - functional dependency rules
    - limitation: dependencies must exist, can not be applied to arbitrary statement
  - **DeFacto** (Lehmann, 2012)
    - applied BOA framework to verbalise the RDF patterns from a training set. Then these patterns were used as queries made to search engines for retrieving relevant webpages and find proofs in the webpages
    - limitation: The training of BOA focused on object properties, no support of datatype properties

Approach

- **Measuring consensus** based on evidence triples collected from other linked datasets
- Following **owl:sameAs** links to collect triples describing same entities in other sources
- **Semantic relatedness** based predicate matching
- **Quantifying** the agreement among sources as an aggregated confidence score
- Compared to other approaches
  - no need to rely on dependency between values of properties (SWIQA)
  - Adequate support for datatype properties (complements DeFacto)

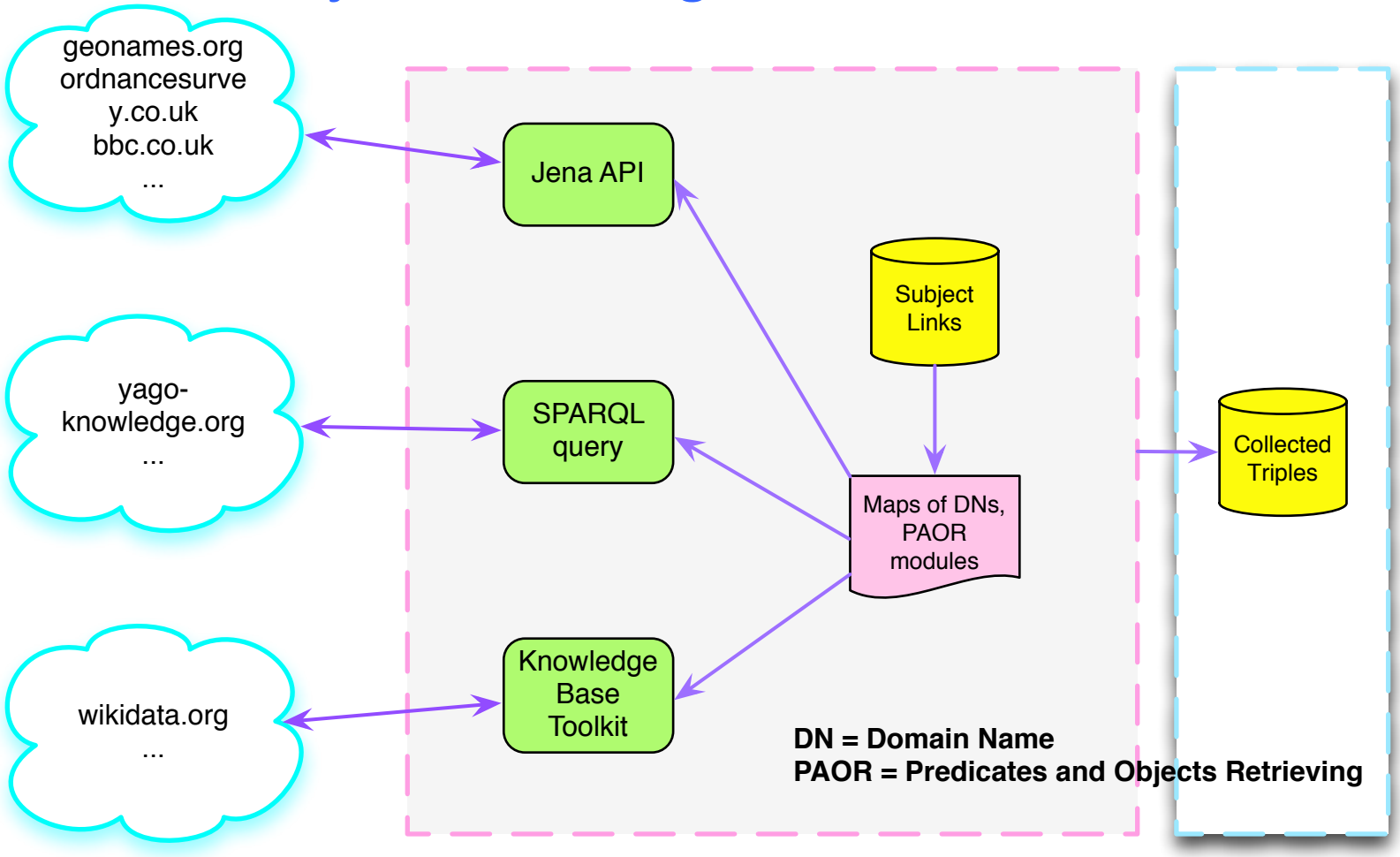
Overview


sameas.org Service
owl:sameAs
Subject Links
1. Collecting & Cleaning
owl:sameAs
Target Data Set
Collect from
Target Triples
Data Set 1
2. Predicates & Objects Retrieving
Data Set 2
Collected Triples
3. Predicate Matching
4. Confidence Measuring
Data Set 3
Confidence of target triples
Number of Triples
10
0
75
50
25
0
Confidence score

Subject Links Collecting and Cleaning

- Fetching equivalent subject links
  - Via the property **owl:sameAs** of the target fact triple
  - Via querying the http://sameas.org service
- Cleaning duplicated and non-resolvable subject links
  - By **pinging** the corresponding URIs
  - Removing **identical URIs**
  - Keeping **English** version of the same resource

Predicates and Objects Retrieving



Predicate Matching

- Semantic relatedness between predicates
  - Not based on string similarity
  - Different strings may have the same meaning (**dbpedia-owl:populationTotal** and **yago:hasNumberOfPeople**)
- The **WuP** [7] semantic relatedness measure is applied
  - considers the depths of the two synsets in the WordNet taxonomies
- The **WS4J** API is used to generate the pairwise similarity matrix for two input predicates (consisting of split compound words)

Predicate Matching

	has	Number	of	People
population	0.0	0.4286	0.0	0.9091
Total	0.0	1.0	0.0	0.3636

1. Pairwise semantic similarity matrix

2. Word Semantic Similarity

$$W(n) = \begin{cases} \max\left(S_{row}(n)\right) & \text{if } r \leq c \\ \max\left(S_{column}(n)\right) & \text{if } r > c \end{cases}$$

$$P = \frac{\sum_{W \in \Phi(W)} W}{k} \text{ with } \exists\ W \in \Phi(W) \text{ and } W > \theta$$

3. Predicate Similarity

Confidence Measuring

- Assign **weighting factors** to subject links to represent their reliability (based on **provenance info**)
- If the object of the target triple is string
  - **Aggregated score**: weighted average of string similarity between the object values of evidence triples and the target triple
- If the object of the target triple is numerical
  - **Subtracting** the **ratio** between difference in object values between target triple and the weighted average score and weighted avg score **from 1**

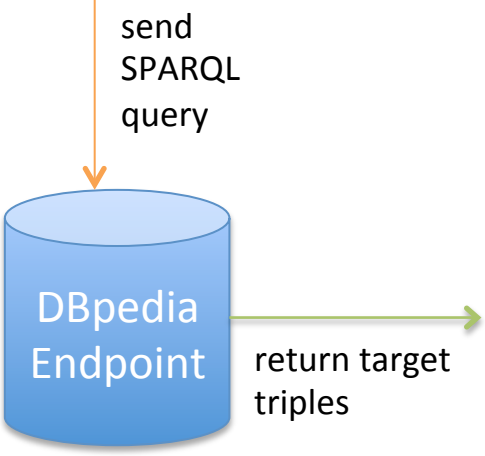
$$\gamma = \frac{\sum_{i=1}^{m} \omega_i \cdot \nu_i}{\sum_{j=1}^{m} \omega_j}$$

$$\Gamma = 1 - \frac{\sqrt{(x-\gamma)^2}}{\gamma}$$

Experiment

- establish the DBpedia test set

```
SELECT ?s dbpedia-owl:populationTotal ?o
WHERE {?s rdf:type dbpedia-owl:PopulatedPlace . ?s dbpprop:postcode-
Area "MK"@en . OPTIONAL {?s dbpedia-owl:populationTotal ?o .}
FILTER (bound(?o)) FILTER (?o > 10000)}
```

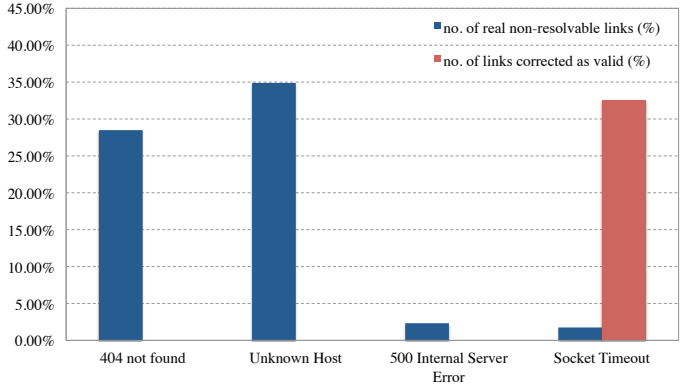



```
1  http://dbpedia.org/resource/Bletchley_and_Fenny_Stratford   dbpedia-
   owl:populationTotal   15313
2  http://dbpedia.org/resource/Wolverton_and_Greenleys   dbpedia-owl:populationTotal
   12492
3  http://dbpedia.org/resource/Bedford   dbpedia-owl:populationTotal 80000
4  http://dbpedia.org/resource/Wolverton dbpedia-owl:populationTotal 12492
5  http://dbpedia.org/resource/Shenley_Church_End   dbpedia-owl:populationTotal 12961
6  http://dbpedia.org/resource/West_Bletchley  dbpedia-owl:populationTotal 22213
7  http://dbpedia.org/resource/Milton_Keynes   dbpedia-owl:populationTotal 229941
8  http://dbpedia.org/resource/Campbell_Park   dbpedia-owl:populationTotal 16402
9  http://dbpedia.org/resource/Newport_Pagnell dbpedia-owl:populationTotal 15118
10 http://dbpedia.org/resource/Walton,_Milton_Keynesdbpedia-owl:populationTotal 11923
11 http://dbpedia.org/resource/Flitwick  dbpedia-owl:populationTotal 12700
12 http://dbpedia.org/resource/Kempston  dbpedia-owl:populationTotal 19440
13 http://dbpedia.org/resource/Buckinghamdbpedia-owl:populationTotal 12043
14 http://dbpedia.org/resource/Woughton  dbpedia-owl:populationTotal 13774
15 http://dbpedia.org/resource/Shenley_Brook_End    dbpedia-owl:populationTotal 25828
16 http://dbpedia.org/resource/Stantonbury     dbpedia-owl:populationTotal 10084
17 http://dbpedia.org/resource/Bletchley dbpedia-owl:populationTotal 33950
18 http://dbpedia.org/resource/Great_Linford   dbpedia-owl:populationTotal 19350
```

http://dbpedia.org/sparql

Results

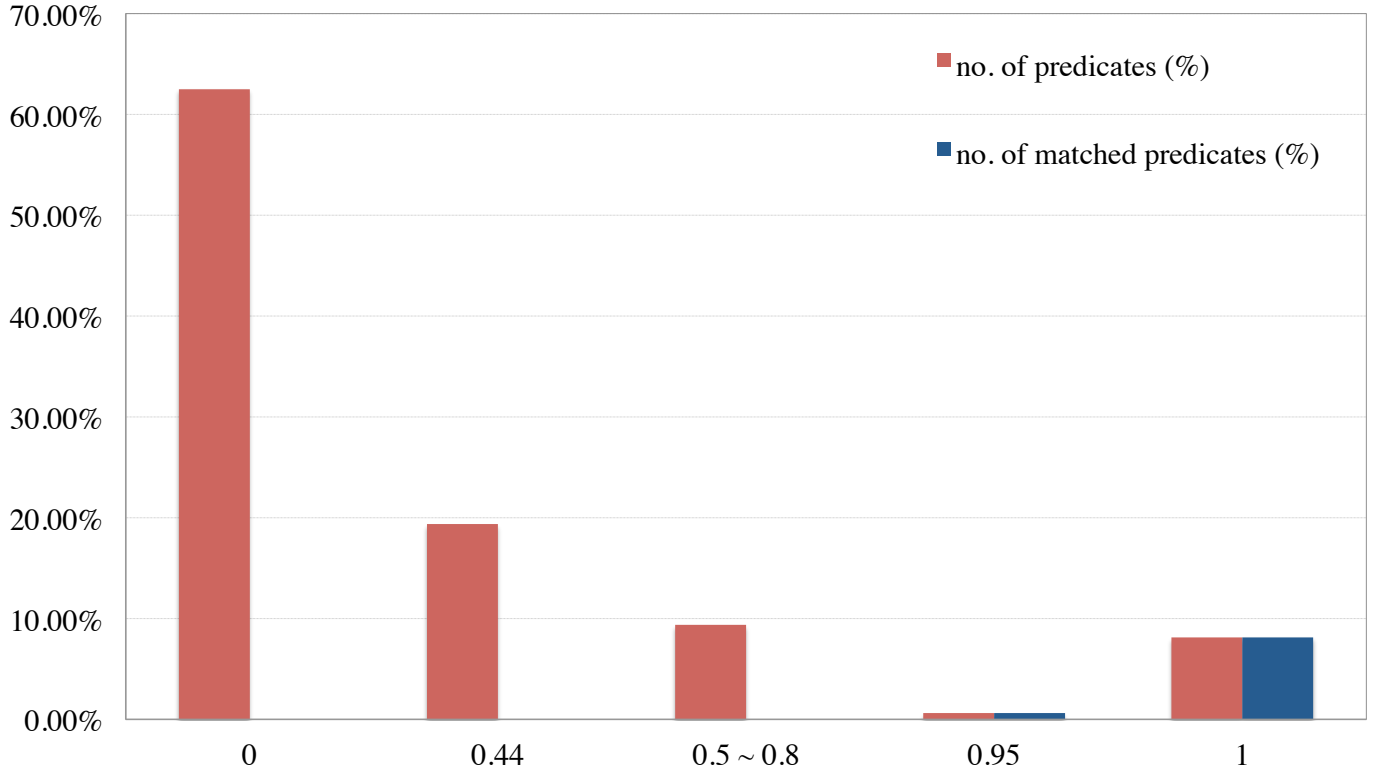
- initially collected **1349** subject links; after the cleaning process, **172** subject links remained
  - **907** non-resolvable links, **18** duplicated links, **252** links in other languages
- improve the cleaning of the subject links process
  - examined the causes of the non-resolvable links



Results

- After the Predicates and Objects retrieving process, **1793** triples collected (for 172 subject links).
- For each target triple, dozens to several hundred triples were collected from other sources.
- Predicate Similarity Distribution
  - **60%** not matched by the algorithm, **40**% matched in different levels by the algorithm

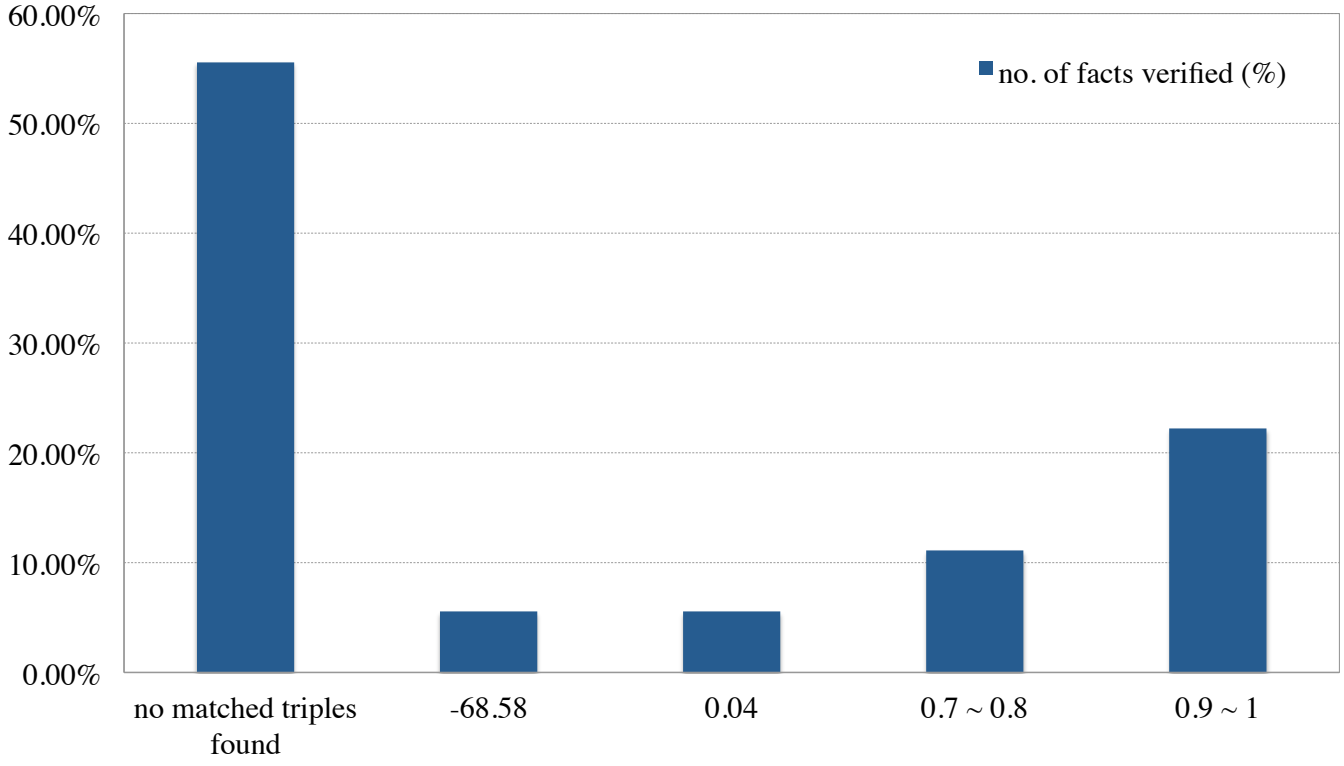
Predicate Similarity Distribution



Results

- Confidence measurement
  - **45%** assigned with a confidence score while **55%** not
  - **22%** of the target triples were identified by the algorithm as highly reliable
- Manually examined the correctness of the confidence scores
  - Low confidence scores due to wrong subject links
  - Need to extend the approach to identify “fake” subject links

Confidence Score Distribution



Conclusion and Future Work

- Our main contributions include a novel approach that enables checking the accuracy of RDF triples or facts.
- The contributions also include identifying matched or relevant RDF triples to a target triple.
- The approach would become increasingly important due to the fast growth of LOD.
- We are planning to demonstrate the approach can be proficiently applied to arbitrary predicates.
- We are also going to evaluate the predicate similarity matching method with standard evaluation measures (Precision/Recall) on well-known datasets.
- We want to explore the correlation between the confidence score assigned with the proposed method and the accuracy of the input fact.

Q & A

Thank You !