QUANT-Question Answering Benchmark Curator

Ria Hari Gusmita, Rricha Jalota, Daniel Vollmers, Jan Reineke, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck

September 10, 2019

Outline

- Mainly based on benchmark datasets (benchmarks)
- Challenge in maintaining high-quality and benchmarks

QALD

LC-QuAD

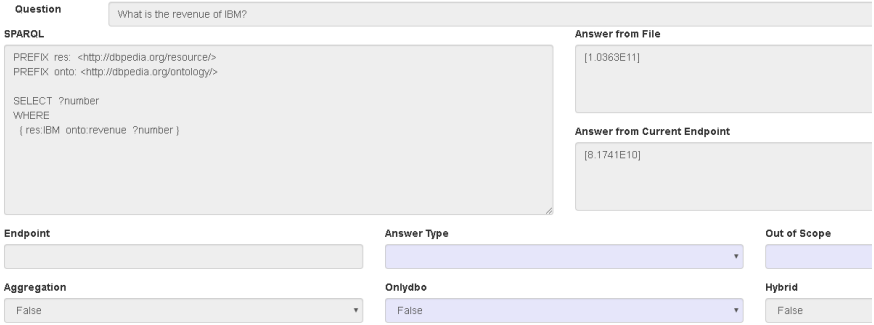
Free917

WebQuestions

- Change of the underlying knowledge base

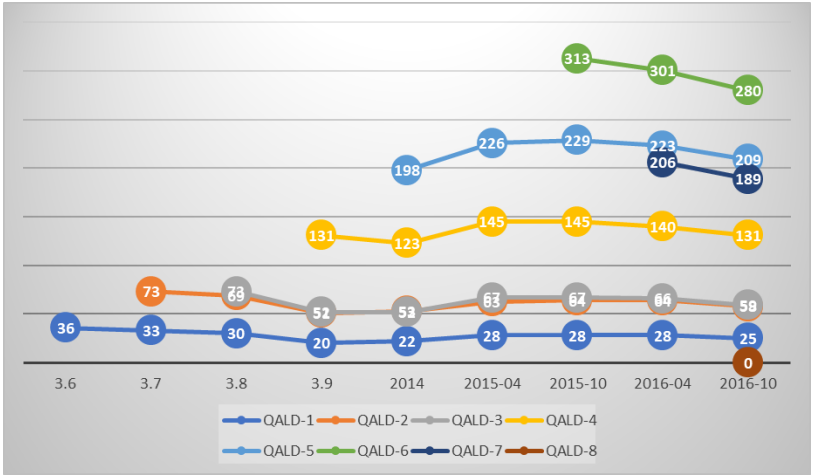
DBpedia 2016-04	DBpedia 2016-10
`http://dbpedia.org/resource/Surfing`	`http://dbpedia.org/resource/Surfer`
`http://dbpedia.org/ontology/seatingCapacity`	`http://dbpedia.org/property/capacity`
`http://dbpedia.org/property/portrayer`	`http://dbpedia.org/ontology/portrayer`
`http://dbpedia.org/property/establishedDate`	http://dbpedia.org/ontology/foundingDate

- Metadata annotation errors



Motivation

Degradation QALD benchmarks against various versions of DBpedia



Contribution

- QUANT, a framework for the intelligent creation and curation of QA benchmarks

Definition

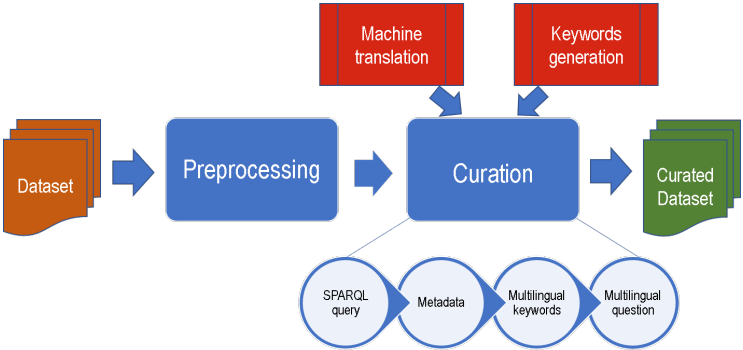
- Given B, D, and Q as benchmark, dataset, and questions respectively
- S represents QUANT's suggestions

  - i^{th} version of a QA benchmark B_i as a pair (D_i, Q_i)
  - Given a query $q_{ij} \in Q_i$ with zero results on D_k with $k > i$
  - $S : q_{ij} \rightarrow q'_{ij}$

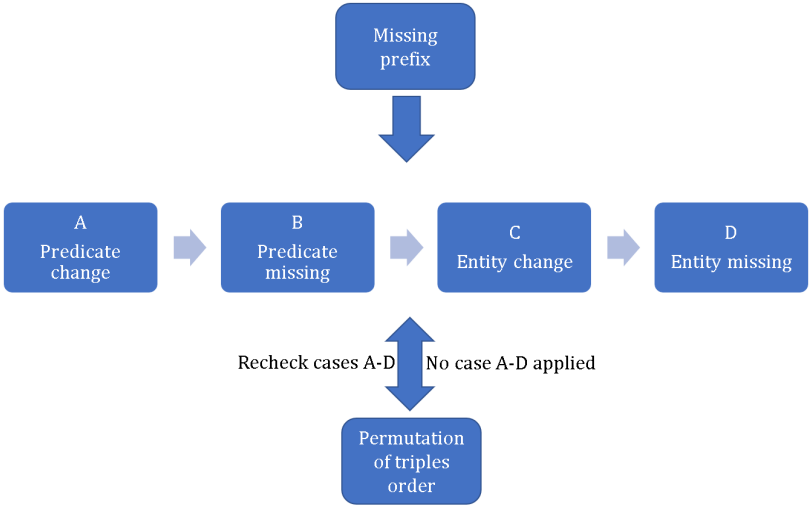
- QUANT aims
  - to ensure that queries from B_i can be reused for B_k
  - to speed up the curation process as compared to the existing one

What QUANT supports

1. Creation of SPARQL queries
2. The validity of benchmark metadata
3. Spelling and grammatical correctness of questions


Machine translation
Keywords generation
Dataset
Preprocessing
Curation
Curated Dataset
SPARQL query
Metadata
Multilingual keywords
Multilingual question

1. SPARQL suggestion
2. Metadata suggestion
3. Multilingual Questions and Keywords Suggestion


Missing prefix
A
Predicate change
B
Predicate missing
C
Entity change
D
Entity missing
Recheck cases A-D
No case A-D applied
Permutation of triples order

- The original SPARQL query

```
SELECT ?s
WHERE {
      res:New_Delhi dbo:country ?s .
}
```

- The original SPARQL query

```
SELECT ?s
WHERE {
    res:New_Delhi dbo:country ?s .
}
```

- The suggested SPARQL query

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
SELECT ?s
WHERE {
    res:New_Delhi dbo:country ?s .
}
```

- The original SPARQL query

```
SELECT ?date
WHERE {
      ?website rdf:type onto:Software .
      ?website onto:releaseDate ?date .
      ?website rdfs:label "DBpedia" .
}
```

- The suggested SPARQL query

```
SELECT ?date
WHERE {
      ?website rdf:type onto:Software .
      ?website rdfs:label "DBpedia" .
      ?website dbp:latestReleaseDate ?date  .
}
```

- The original SPARQL query

```
SELECT ?uri
WHERE {
      ?subject rdfs:label "Tom␣Hanks".
      ?subject foaf:homepage ?uri
}
```

- The original SPARQL query

```
SELECT ?uri
WHERE {
      ?subject rdfs:label "Tom␣Hanks".
      ?subject foaf:homepage ?uri
}
```

- The suggested SPARQL query The predicate foaf:homepage is missing in ?subject foaf:homepage ?uri

- The original SPARQL query

```
SELECT ?uri WHERE
{ ?uri rdf:type yago:CapitalsInEurope }
```

- The original SPARQL query

```
SELECT ?uri WHERE
{ ?uri rdf:type yago:CapitalsInEurope }
```

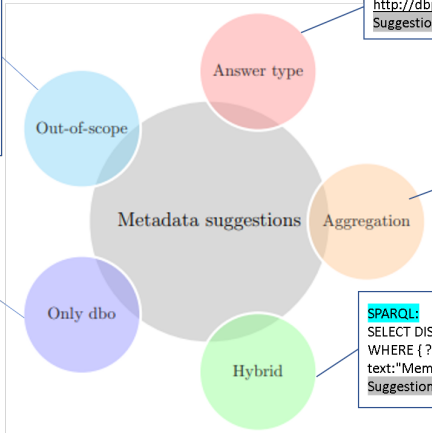
- The suggested SPARQL query

```
SELECT ?uri WHERE
{  ?uri rdf:type yago:WikicatCapitalsInEurope }
```

2. Metadata suggestion

SPARQL:
```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
SELECT DISTINCT ?uri
WHERE { ?uri dbo:occupation res:Surfing ;
dbo:birthPlace res:Australia }
```
Answer from current endpoint:
-
Suggestion: True

SPARQL:
```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
SELECT DISTINCT ?n
WHERE
 { res:The_Lego_Movie
        dbo:budget ?n
 }
```
Suggestion: True



Answer from current endpoint:
http://dbpedia.org/resource/Japan
Suggestion: resource

SPARQL:
```
SELECT COUNT (?medal) as ?c
WHERE {
?medal <http://dbpedia.org/ontology/goldMedalist>
<http://dbpedia.org/resource/Michael_Phelps>. FILTER
regex(lcase(str(?medal)), "olympics") .}
```
Suggestion: True

SPARQL:
```
SELECT DISTINCT ?uri
WHERE { ?x text:"is" text:"sports commentator" . ?x text:"known as"
text:"Memory Man" . ?x <http:/dbpedia.org/ontology/residence> ?uri.
```
Suggestion: True

Question with missing keywords and translations


```
"id": "2",
"datasetVersion": null,
"answerType": "",
"aggregation": false,
"hybrid": false,
"onlydbo": false,
"sparqlQuery": "PREFIX yago: <http://dbpedia.org/class/yago/>\nPREFIX rdf: <http:
http://dbpedia.org/property/>\nSELECT ?uri ?string\nWHERE \n{\n\t?uri rdf:type ya
?string. FILTER (lang(?string) = 'en') }\n} ORDER BY ASC(?density) LIMIT 1",
"pseudoSparqlQuery": null,
"outOfScope": null,
"languageToQuestion": {
    "en": "Which state of the United States of America has the highest density?"
},
"languageToKeyword": {},
"goldenAnswer": [
    "http://dbpedia.org/resource/New_Jersey"
]
```

- Generated keywords: state, united, states, america, highest, density
- Utilizing Trans Shell tool→Generated keywords translations suggestion

```
"languageToKeyword" : {
  "de" : [ "Zustand ", "vereinigt ", "Zustände ", "Amerika ", "höchste ", "Dichte " ],
  "ru" : [ "государство ", "единый ", "состояния ", "Америка ", "наибольший ", "плотность " ],
  "pt" : [ "Estado ", "Unidos ", "estados ", "América ", "maior ", "densidade " ],
  "en" : [ "state ", "united ", "states ", "america ", "highest ", "density " ],
  "hi_IN" : [ "राज्य", "संयुक्त", "राज्यों", "अमेरिका", "उच्चतम", "घनत्व" ],
  "it" : [ "stato ", "unito ", "stati ", "America ", "massimo ", "densità " ],
  "fa" : [ "ایالت ", "متحد ", "ایالات ", "آمریکا ", "بیشترین ", "تراکم " ],
  "fr" : [ "Etat ", "uni ", "États ", "Amérique ", "le plus élevé ", "densité " ],
  "ro" : [ "stat ", "Unit ", "statele ", "America ", "cel mai inalt ", "densitate " ],
  "es" : [ "estado ", "unido ", "estados ", "America ", "más alto ", "densidad " ],
  "nl" : [ "staat ", "verenigd ", "staten ", "Amerika ", "hoogst ", "dichtheid " ]
},
```

Suggested Question Translations

```
"languageToQuestion": {
    "de": "Welche Zustand von das Vereinigt Zustände von Amerika hat das höchste Dichte? ",
    "ru": "Который государство из  объединенный состояния из Америка имеет  наибольший плотность? ",
    "pt": "Qual Estado do a Unidos Estados do América tem a maior densidade? ",
    "en": "Which state of the United States of America has the highest density?",
    "hi_IN": "कौन कौन से राज्य का यूनाइटेड राज्य अमेरिका का अमेरिका है उच्चतम घनत्व?",
    "it": "Quale stato di il Unito stati di America ha il massimo densità? ",
    "fa": "مکارت؟ نیرتالاب نیا اکیرمآ یاراد نیا زا تلایا اه دتیانوی نیا زا تلاج مادک ",
    "fr": "Lequel Etat de la Uni États de Amérique a la le plus élevé densité? ",
    "ro": "Care stat de  Unit statele de America are  cel mai inalt densitate? ",
    "es": "Cual estado de el Unido Estados de America tiene el más alto ¿densidad? ",
    "nl": "Welke staat van de Verenigd Staten van Amerika heeft de hoogst dichtheid? "
},
```

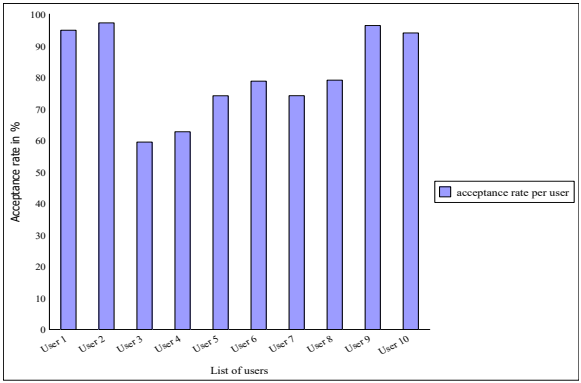
Evaluation

- Three goals of the evaluation:
  1. QUANT vs manual curation
     - Graduate students curated 50 questions using QUANT and another 50-question manually
     - 23 minutes vs 278 minutes
  2. Effectiveness of smart suggestions
     - 10 expert users got involved in creating a new joint benchmark, called QALD-9, with 653 questions
  3. QUANT's capability to provide a high-quality benchmark dataset
     - The inter-rater agreement between each two users amounts up to 0.83 on average

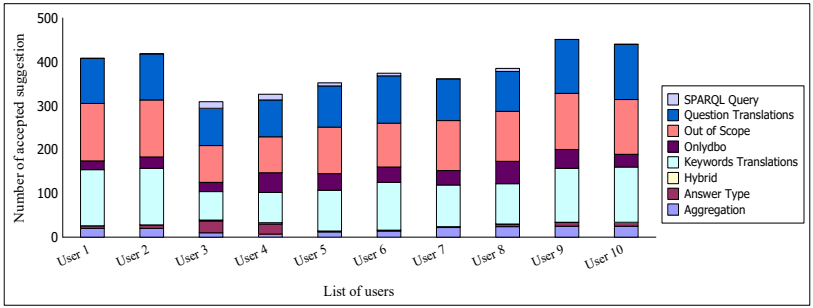
Group	Inter-rater Agreement
1st Two-Users	0.97
2nd Two-Users	0.72
3rd Two-Users	0.88
4th Two-Users	0.77
5th Two-Users	0.96
Average	0.83

Evaluation

Users acceptance rate in %



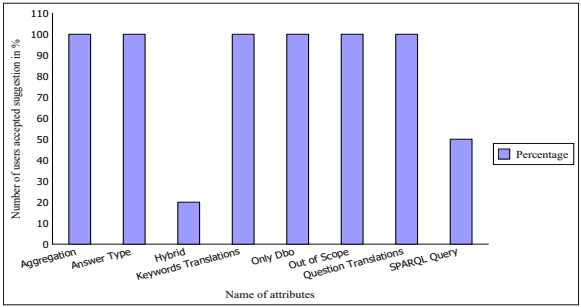
- QUANT provided 2380 suggestions and user acceptance rate on average is 81%
- The top 4 acceptance-rate are for QALD-7 and QALD-8



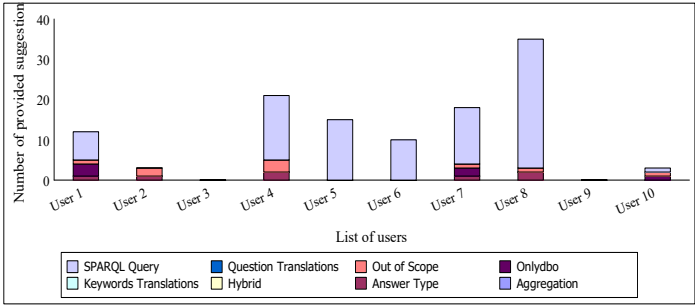
- Most users accepted suggestion for out-of-scope metadata
- Keyword and question translation suggestions yielded the second and third highest acceptance rates.

Evaluation

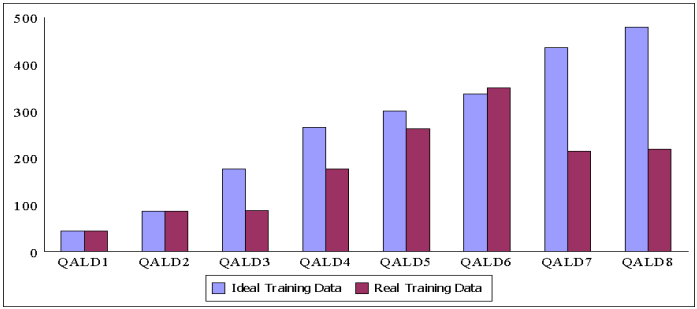
Number of users who accepted QUANT's suggestions for each question's attribute.



- 83.75% of the users accepted QUANT's smart suggestions on average
- Hybrid and SPARQL suggestions were only accepted by 2 and 5 users respectively.



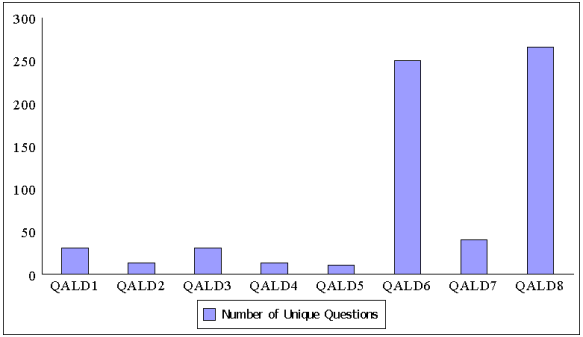
- Answer type, onlydbo, out-of-scope, and SPARQL query metadata were attributes whose value redefined by users



There are 1924 questions where 1442 questions are training data and 482 questions are test data

QALD-specific Analysis

- Duplication removal resulted 655 unique questions
- Removing 2 semantically similar questions produced 653 questions
- Using QUANT with 10 expert users, we got 558 total benchmark questions → increase QALD-8 size by 110.6%
- The new benchmark formed QALD-9 dataset



Distribution of unique questions in all QALD versions

Conclusion

- QUANT's evaluation highlights the need for better datasets and their maintenance
- QUANT speeds up the curation process by up to 91%.
- Smart suggestions motivate users to engage in more attribute corrections than if there were no hints

- There is a need to invest more time into SPARQL suggestions as only 5 users accepted them
- We plan to support more file formats based on our internal library

Thank you for your attention!

Ria Hari Gusmita
ria.hari.gusmita@uni-paderborn.de
https://github.com/dice-group/QUANT
DICE Group at Paderborn University
https://dice-research.org/team/profiles/gusmita/