FRanCo – A Ground Truth Corpus for Fact Ranking Evaluation

Tamara Bobić, Jörg Waitelonis, and Harald Sack

Hasso-Plattner-Institute,
Prof.-Dr.-Helmert-Str. 2-3,
14482 Potsdam, Germany
{tamara.bobic,joerg.waitelonis,harald.sack}@hpi.de

Abstract. The vast amount of information on the Web poses a challenge when trying to identify the most important facts. Many fact ranking algorithms have emerged, however, thus far there is a lack of a general domain, objective gold standard that would serve as an evaluation benchmark for comparing such systems. We present *FRanCo*, a ground truth for fact ranking acquired using crowdsourcing. The corpus is built on a representative DBpedia sample of 541 entities and made freely available. We have published both the aggregated and the raw data collected, including identified nonsense statements that contribute to improving data quality in DBpedia.

Keywords: fact ranking, corpus acquisition, crowdsourcing

1 Introduction

Since its early days, the Semantic Web community has focused on turning the unstructured web into a structured "web of data" [14]. The Linked Open Data (LOD) project interlinked diverse sources of information and gave rise to the world's largest publicly available knowledge base, currently comprising more than 74 billion facts[1] [2]. DBpedia, a large-scale knowledge base extracted from Wikipedia, is the most interlinked Dataset of the decentralized LOD [11].

The sheer amount of information in DBpedia alone imposes a challenge when presenting entities and their properties in a concise form to the human user, (*e. g.* LOD visualization) or via LOD mashups. The English version of the DBpedia 2014 data set currently describes 4.58 million entities with 583 million facts[2] in the form of RDF triples. Thereby, on average, each entity is described by 127 facts. These facts are not ordered or ranked in any way, making it unclear which of them are important and should be included in a concise representation of the entity.

This overflow of information gave rise to fact ranking, which is a crucial step in deciding which statements are most relevant and informative for describing

[1] As of August 2014, `http://lod-cloud.net`

[2] `http://wiki.dbpedia.org/Datasets`

an entity. The relevance of facts undoubtedly depends on the context and the user's needs. It might seem obvious that *"Slovenia is a country in Europe"* is more important than *"Slovenia has 0.7% of water area."*. However, these facts could be disparately ranked among different users and purposes. Nevertheless, after taking into account the multitude of possible contexts and usages, we have decided to focus on a general information need, which considers the average human view.

The major web search engines have recognized the need for fact ranking and summarization of their search results. The most prominent one, Google Knowledge Graph, produces structured and salient summaries of entities, using some of the available Linked Data knowledge bases [15]. In recent work they also adapt their model to account for trustworthiness and relevance of facts contained in a web page [4]. Furthermore, we have seen much effort in the direction of fact ranking and entity summarization [8, 10, 15, 18] (discussed in Sect. 2). Many of these approaches lacked a comparative benchmark with other systems, due to a nonexistent generic and comprehensive gold standard. Thus far, several efforts have gone towards the creation of manually curated ground truths, but have fallen short to provide: objectivity (annotated by a small user sample, usually from the same location [10]), generalizibility (focused on just one domain, *e. g.* persons, movies [17]) and significant corpus size (usually too small [7, 10, 17]).

The contribution of this paper is the generation of *FRanCo*, a ground truth dataset that enables a generic and standardized quantitative evaluation of fact ranking systems. Following a crowdsourcing approach, we have generated a corpus that includes opinions of hundreds of users about a diverse subset of DBpedia entities, providing a more objective and comprehensive insight into the relevance of DBpedia facts. We have used a semi-supervised approach to generate a representative sample of DBpedia entities and propose a method to calculate a ground truth ranking of facts based on the opinions provided by the users. The corpus is made publicly available in RDF format[3] and can be used as a building block for the development of novel ranking and summarization techniques on Linked Data.

The remainder of this paper is structured as follows: Sect. 2 gives an overview of the previous work regarding fact ranking strategies, as well as similar attempts for corpus generation. We further introduce our effort for corpus acquisition in Sect. 3, providing more details about the chosen DBpedia sample, user interface, and obtained data statistics. Sect. 4 presents the dataset structure and ranking measurements. Finally, in Sect. 5 we conclude and list future efforts to maintain a high data quality of the corpus.

2 Related Work

The need to rank associations extends into several research fields and is valuable whenever there is a need for summarization or prioritization of such. Numerous

[3] `http://s16a.org/node/13`

areas that can be explored using graph traversal algorithms, such as recommender systems or named entity processing, can directly benefit from ranking heuristics. Exploratory systems based on Linked Data enable the discovery of new associations among resources and assist the user in exploring the data space. E. g., in [21] the authors present a system for exploratory video search which employs several heuristics for ranking properties of LOD resources, such that the most relevant associations are used for the generation of further search suggestions.

Algorithms which exploit the structural aspects of Linked Data graphs are in principle a good choice for the ranking of RDF resources. Many ranking systems have been adaptations of well established and scalable algorithms like PageRank [8, 9, 18] or HITS [1, 5]. However, the semantic layer of RDF knowledge bases is usually neglected in these approaches. Links often are of different type, meaning and relevance, which is not exploited by these algorithms. The ReConRank [8] algorithm relies on PageRank to compute the relevance of resources (ResourceRank), but in addition also exploits the context of a specifc resource (ContextRank). RELIN [3], an entity summarization system, modifies the random surfer model to favor related and informative measures. It provides a summary of limited size, with the goal to select distinctive information which identify an entity, but are not necessarily important. On the other hand, DIVERSUM [16] and FACES [6] aim to provide diversity, along with important characteristics of an entity. They give preference to variety over relevance, in order to reduce redundancy in the result set. TripleRank [5] extends the HITS algorithm by applying a decomposition of a 3-dimensional tensor that represents an RDF graph. Its approach provides faceted authority ranking results and also enables navigation with respect to identified topics. A similar work by [1] that computes "subjective" and "objective" metrics, which correspond to hubs and authorities is also based on a HITS type architecture. Many of the ranking systems perform an intrinsic evaluation, judging the output in comparison to a gold standard result, as pre-defined by a small number of human evaluators. However, such evaluations are rarely reproducible and don't offer a standardized comparison to other ranking heuristics. Due to a rising trend and an overwhelming number of emerging ranking systems, several efforts have been made to construct a reference dataset that would serve as general ground truth.

Creation of gold standard datasets can be a strenuous, time consuming and expensive task. An attempt to overcome these difficulties is a silver standard benchmark like DBpediaNYD [12] – a large-scale, automatically created dataset that contains semantic relatedness metrics for DBpedia resources. The rankings are derived from symmetric and asymmetric similarity values, as measured by a web search engine. However, being machine generated, the corpus should be used with caution and only as a complement to manually generated gold standards. Identifying a "general truth" is an inherently subjective problem which cannot yet be reliably solved automatically. Maintaining the focus on DBpedia as a comprehensive and general knowledge base, [10] analyzes various strategies for fact ranking. For evaluation purposes, a two-fold user study was conducted which resulted in a reference dataset that can potentially be used for comparison of different ranking heuristics. However, this dataset is rather small, covering only 28 DBpedia entities, as evaluated by 10 human judges and is not available at the published location. The advantages of crowdsourcing over expert-annotated data is the access to a "wider market" of cultures and languages. However, gathering user opinions through crowdsourcing may turn out to be challenging when it comes to attracting and motivating the users. Games with a purpose have emerged as a platform that mitigates this drawback by incorporating the element of fun into the process of knowledge harvesting.

WhoKnows? [20] is an online quiz game with the purpose of gathering opinions about relevant LOD properties, which would in turn serve for crafting more refined heuristics for semantic relatedness of entities. It was initially designed to evaluate ranking heuristics proposed in [21]. However, the gathered data has not been made available in form of a fact ranking dataset. *WhoKnows?Movies!* [17] is designed in the style of *"Who Wants to Be a Millionaire?"*, presenting multiple choice questions to players. The relevance of an individual property (fact) is determined as a function of its popularity among the game players. The chosen sample consists of 60 movies taken from the IMDb[4] Top 250 list. After obtaining inputs from 217 players who played 690 times, the authors provide an evaluation of the UBES system [19] and GKG [15] on their dataset. The created fact ranking gold standard was made publicly available, however, its relatively small size and restriction to the narrow context of movies, raises the question of generalizability. BetterRelations [7] is a two player agreement game, where in each game players are presented with an entity (topic) and two facts that describe it. Players are then supposed to decide which of the facts is more important, while also having the option to skip or report both facts as nonsense. Fact ratings are updated after each atomic competition, minimizing the number of decisions needed. The sample consisted of 12 DBpedia topics covering diverse domains and the game was played 1041 times by 359 users. However, to the best of our knowledge, the obtained dataset is not publicly available.

Overall, we have observed a lack of a publicly available, generic and objective datasets which would serve as a benchmark. We address this issue and present our crowdsourcing effort to collect the knowledge of people globally, covering a wide range of topics in a publicly available, high quality gold standard dataset.

3 Ground Truth Generation

Following the crowdsourcing approach, we have designed a user interface to derive the relevance of facts from the opinions of many. In this section we further present the process of selecting a representative DBpedia sample, the user interface and its interaction entities, as well as the statistics obtained so far.

[4] `http://imdb.com/`

3.1 DBpedia Sample

The Linked Data counterpart of Wikipedia, DBpedia, represents a general encyclopedia, covering a large variety of topics and entities. We acknowledge that DBpedia does not necessarily encompass the entire human knowledge, however, it does offer a representative snapshot of collective knowledge, since it is being created by people from all around the world.

In order to draw a representative sample which covers different domains of knowledge (*e. g.* persons, places, creative works etc.), we look at the underlining semantic structure provided in the DBpedia ontology. This is a shallow, cross-domain ontology, which has been manually created based on the most commonly used infoboxes within Wikipedia[5]. At the time of the sample creation, the ontology covers 382 classes.

We have chosen not to base the sample only on the most popular entities, but try to cover the broad landscape of human knowledge present in DBpedia. In order to reduce redundancy and maintain high quality of the data, we have selected only the RDF triples that contain properties mapped to the DBpedia ontology (`http://dbpedia.org/ontology/`). We have ignored technical, structural, and administrative properties which are not useful in the fact ranking scenario (*e. g. abstract, imageSize, thumbnail*) and focus on descriptive properties. Additionally, we have included triples with the `dcterms:subject` property, which we consider essential, since they denote the Wikipedia categories a particular entity is part of.

When considering all the subject-property-object triples that describe an entity, we have included both the direct (entity is the subject) and the inverse (entity is the object) ones. We have found inverse triples to be highly valuable when describing an entity, since many entities (*e. g. Pacific Ocean, Africa, Microsoft Windows, English people, Heart*) have the majority of the information encoded in the inverse direction.

Automatic Pruning Phase In the automatic pruning phase, we have filtered DBpedia ontology classes based on the number of entities a class contains, their popularity and the position of the class in the hierarchy tree of the ontology. An ontology class was not considered if it is:

1. Too general
   (a) has more than 120000 members
   (b) has 3 or more hierarchical levels underneath
2. Too specific and insignificant
   (a) too low in the tree hierarchy (4th level or below)
   (b) the max inDegree of its entities is < 500

We have chosen the threshold of 120000 members since anything lower would aggressively prune out many important (*i. e.* popular) classes. Classes like *Agent,*

[5] `http://wiki.dbpedia.org/Ontology`

Person, Place, Organization etc. are eliminated in step 1a). Many classes have an elaborate hierarchical structure underneath them which indicates that the class is very abstract and probably can be disregarded, *e. g. Event, Region, Activity, NaturalPlace, WrittenWork*. Similarly, if a class is too low in the hierarchy, we assume that it might be too specific and hence not essential, *e. g. SnookerChamp, RailwayLine, Manga*. In addition we take into account entity popularity in terms of the in-degree of their link graph, as defined by `dbpedia-owl:wikiPageInLinkCount`. We have chosen 500 as the minimum threshold for a class to be included in the sample. Some of the classes eliminated in this step are *Lymph, Locomotive, SpaceShuttle*. Finally, there are several cases of almost identical classes, *i. e.* the members of the child class are almost the same as in the parent class (*e. g. SoccerManager* and *SportsManager*, *LegalCase* and *Case*, *BiologicalDatabase* and *Database*). After eliminating all of the classes that fall under the above specified criteria, we have reduced the number of considered classes from 382 to 189.

Manual Pruning Phase The purpose of the final sample is to represent the broad spectrum of topics in DBpedia, while at the same time including only well known entities. Therefore, from each of the pre-filtered 189 DBpedia classes, we have selected the top 10 entities based on their in-degree, thus giving priority to those which are most popular. However, the number of links referring to an entity, does not necessarily denote its real life recognizability. Entities like *Bofors 40 mm gun* (class *Weapon*), *Plesetsk Cosmodrome* (class *MilitaryStructure*), *Brown v. Board of Education* (class *Case*) seem rather obscure and might not be known to many people, despite their high in-degree values.

In order to leverage the representativeness of the sample by its recognizibility to the user, we have carried out a manual evaluation step. For each of the 1890 entities, 8 human judges have provided their personal opinion on whether the entity should be included as an 'important' entity in the final sample. The judges vote positively if they are familiar with the entity and find it important, considering also that people from different cultural backgrounds might share the same view.

After merging opinions, we have shortlisted those entities with at least 3 votes, which resulted in a sample of size 574. Out of 1890, 547 entities received 0 votes, *i. e.* has been deemed as unimportant by all the judges, while on the other side 31 entities have been considered important by all, such as *e. g. Coca Cola, Albert Einstein, and NASA*.

Final Sample To reduce the noise we have eliminated facts that provide no additional information about the entity, *e. g. "Albert Einstein is a member of category Albert Einstein"*. For the same purpose, we have constructed a list of 17 symmetric (*e. g. spouse, relative*) and inverse properties (*e. g. child-parent, predecessor-successor*) and removed all duplicate facts such as *"Kant influenced Chomsky" – "Chomsky was influenced by Kant"*. Multivalued literals were present in 116 out of 541 entities and sometimes were redundant or have contained erroneous information. Additionally, there were cases when several object type properties point to the same information (*e. g.* Danube has Black Mountain as *sourceMountain, sourcePlace* and *sourceRegion*). However, we have chosen not to address these issues and kept all the existing values, as any manipulation would introduce bias. Given a large statistical sample, these irregularities will cancel themselves out and the most important facts will emerge to the surface.

We further have decided to eliminate all entities that obtain more than 150 facts (*e. g. United States, Winston Churchill*), so as not to overwhelm the users of the fact ranking tool with too much information. Additionally, we have kept only the entities with at least 10 represented facts (*e. g. Heart, Mona Lisa*), since a smaller number might impede the fact ranking process. The final sample contains 541 entities and 26198 triples. On average, an entity has 48 facts with 15 different properties. 522 entities have inverse properties. Overall, the ratio of inverse properties in the sample is 40.7%.

3.2 User Interface

Following the best-practice guidelines for corpus acquisition using crowdsourcing [13], we have kept the task simple and intuitive for the users with a clean undistractive interface. Upon registering, users are asked to fill out a short survey for statistical purposes. Fig. 1 shows the two steps of the user interface[6]. The entities presented to users are chosen randomly.

In Step 1, the user is asked to list the most relevant facts that come to mind, given an entity and its photo. The answers should be provided as keywords. This part is based on the personal knowledge of the users. Additionally, the users specify the level of confidence they have for the entity at hand, ranging from 1 to 4 (*very familiar, fairly familiar, slightly familiar, not familiar*).

Step 2 list all the facts of an entity, in the form of natural language sentences automatically generated from RDF triples (*e. g. "CNN has headquarter Atlanta"*). For reasons of usability and efficiency, the facts are grouped by properties and limited to 10 per page. Using the radio buttons, users select the importance of each fact on the Likert scale, scoring it from 1 to 5 (high to low). Additionally, the *"I don't know"* button (selected by default) can be checked if the user is unsure or not familiar with the fact, whereas the *"nonsense"* button serves to report a faulty or nonsensical triple. By providing answers the user can score points. While working on the task, the users are given information about the number of entities completed, their current score and the potential gain of completing the current page. Users can resume an interrupted session at their convenience.

In order to incentivise the users by appealing to their competitivness, we have devised a scoring heuristic which reflects users' contribution and efficiency. The number of points earned is a function of the number of facts for which the importance was chosen. If all of the facts were ignored while the user's confidence was at least "slightly familiar", a penalty is assigned. Furthermore, we account for cheating by penalizing users that are working too fast, *i. e.* did not take sufficient time to adequately read the facts. The minimum amount of time needed

[6] `http://s16a.org/fr/`

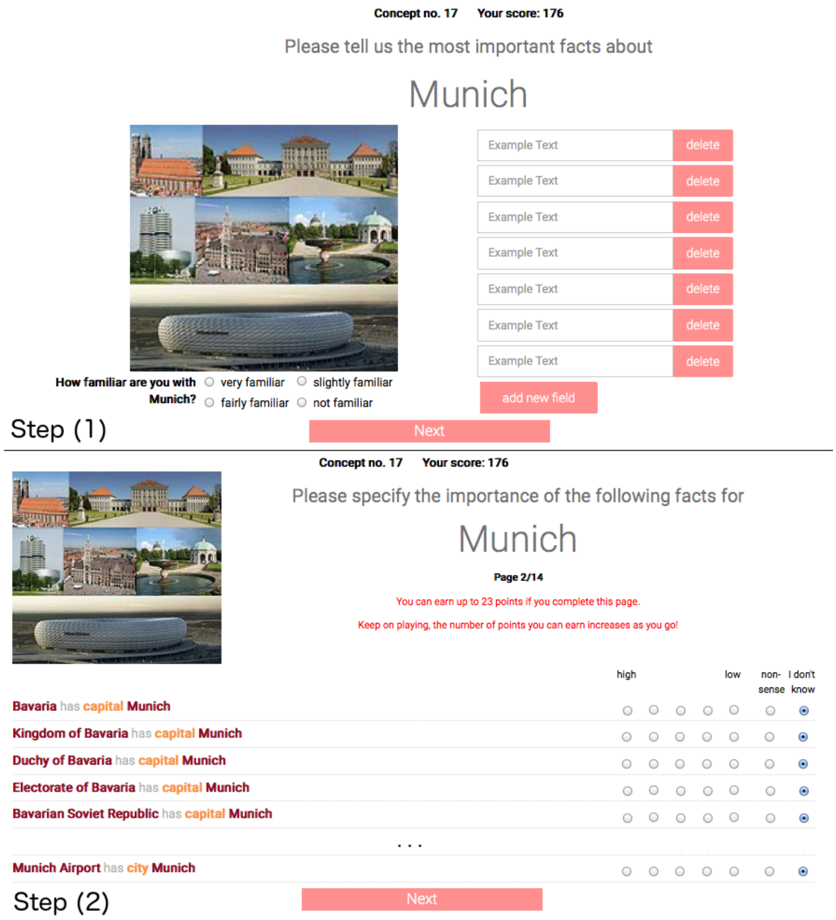
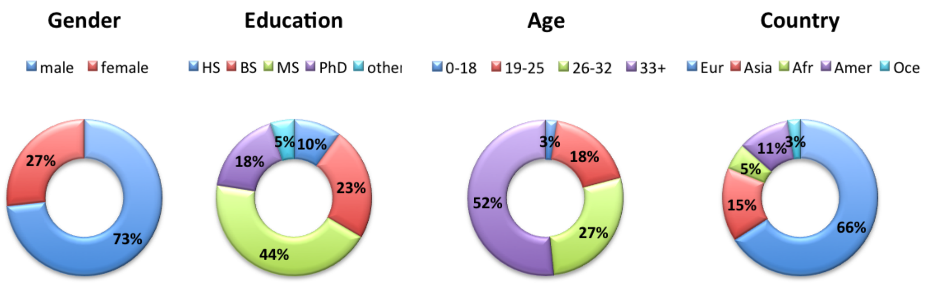


Fig. 1. User interface of the ground truth generation tool.

is determined based on the number of characters in the sentences and the tested average reading speed of 8 human judges. We award extra points if the users help us to identify nonsense statements. The user's score, together with the top scoring participants are presented in a high score table.

3.3 Fact Ranking Statistics

So far, the fact ranking tool has been used by 570 participants, who have covered 3606 entities in total (on average a user has worked on 6.33 entities). Fig. 2 gives an overview of the users' demographics, in terms of their gender, age, education and country of origin. The application has attracted more males than females and has a rather uniform distribution over the age groups. In terms of education, most users hold or pursue a Masters degree, followed by a Bachelors degree, PhD, high school and other. The dominance of people with higher education is due to our advertising on many academic channels and mailing lists. Additionally, we have collected information about a user's country of origin and country of current residence, since in many cases these two differ, which may impact the user's knowledge scope and opinions due to the cultural transfer. A wide audience has been reached, having users from 82 different countries, most of them being in Europe, followed by Asia, the Americas, Africa and Oceania.



(a) Gender (b) Education (c) Age (d) Country of origin

Fig. 2. Demographic overview of the participants.

Given that at least 5 users provided input, the average confidence of the users for the processed entities is 2.70 (on a 1–4 scale, 1 being the most confident), which falls in the middle range. *XML* is the entity with the highest confidence (1.17), while *One Piece* (Japanese Manga series) is most unknown to users (3.83). Although the entities have been assigned to users in a randomized order, some of them have been processed by more users than others. Out of the 541 entities of our sample, 265 entities have been processed by 5 or more users. Fig. 3 shows the distribution of users over entities. On average, there are 4.45 users per entity.

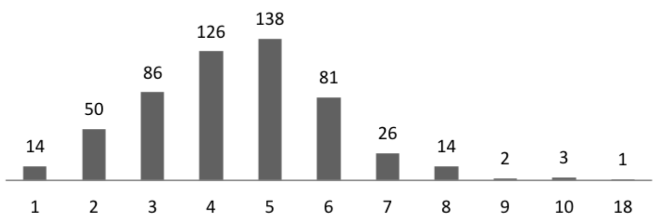


Fig. 3. Distribution of participants over entities. The Y axis represents the number of entities which have been processed by a specific number of participants (X axis).

Our assessment showed an overall fair agreement among the users, with an average Cohen's kappa value of 0.39. The highest agreement reached was 0.90, while some entities had a negative kappa score, which greatly affected the final average value. We considered the facts marked as *"nonsense"* to have a low importance, assuming that users regard nonsensical statements as erroneous and not appropriate for describing the entity at hand. In addition, we assume that facts which are unfamiliar to users (no assigned score, marked as *"I don't know"*) have a low importance. Due to the ordinal nature of the ratings, linear weights were used in the calculation.

The average number of textual inputs in Step 1 is 3. Users' confidence about the entities is moderately correlated with the number of answers provided in this phase, as indicated by the Pearson's correlation coefficient of 0.36.

The data collected in Step 2 reveals 2201 unique nonsense statements, such as *"Dirk Nowitzki has birth place Basketball Bundesliga"* or *"BlackBerry 10 has programming language Java"*, where Java refers to the island. Moreover, we have calculated the standard deviation of users' opinions and have identified the facts whose importance was most and least agreed upon. For example, the majority has agreed that *"Apollo 11 has crew member Neil Armstrong"* is highly important, however, there was a significant disagreement whether the movie *Full Love* is highly important when describing *Jean-Claude Van Damme*. These disputed statements can be indicative of controversial facts, diversity of opinions, but also lack of clarity and alternative interpretations of the presented sentences.

4 *FRanCo* (Fact Ranking Corpus)

FRanCo is generated by aggregating the inputs from our fact ranking tool. In the following, we present our statistics for calculating the ground truth ranking and the structure of *FRanCo*.

4.1 Fact Rank Calculation

Given a vector of users' inputs, we have aggregated the values on fact level, in order to calculate the average score which captures the importance of a fact. In addition, we have formulated a weighted average score, based on the following assumptions:

1. The higher the confidence of the user, the more relevant the answers.
2. The less people are familiar with the fact, the less important it is.
3. The more people know about a fact, the more important it is.

The first assumption pertains to the self-reported confidence of users. Answers of users who are "highly confident" are weighted as 3 fold, "fairly confident" as 2 fold, "slightly confident" as 1 fold and "not confident" as 0.5 fold. The second assumption is related to the number of times the *"I don't know"* button has been checked. If the majority of users are not familiar with the fact, we regard it as less important and penalize its score. Finally, we have analyzed the user inputs from Step 1 and when possible, matched them to facts of the entity under consideration. These inputs represent the personal knowledge and indicate facts that are regarded as important by the user. The higher the frequency of a specific input, the more the importance of its corresponding fact is increased.

In addition to the average and weighted average scores, we also report their values normalized to the interval $[0, 1]$.

4.2 Published Dataset

The collected data is published at `http://s16a.org/node/13` and consists of two main parts: aggregated statistics in RDF format and anonymized raw data. The aggregated corpus is the core of *FRanCo*, primarily intended for evaluation of fact ranking algorithms. For each entity from the DBpedia sample, it contains the number of users, list of facts with ranks, standard deviation of opinions for each fact and the number of times a fact has been reported as nonsense. To measure the performance of ranking algorithms and their closeness to the ground truth, we recommend Kendall's τ and Spearman's ρ, established information retrieval measurements for comparing ranked lists. Additionally, we propose the use of Discounted Cumulative Gain and a more recent Rank Biased Overlap [22], two top-weighted approaches which give more importance to items at a higher rank. The anonymized raw data contains all the information gathered from the fact ranking tool, including user profiles, the DBpedia sample facts and user inputs from Step 1 and Step 2. Additionally, for Step 1 we provide the mapping of user inputs to DBpedia entities achieved with our in-house named entity mapping system KEA[7].

5 Conclusion

In this paper we have presented a crowdsourcing approach for generating *FRanCo*, a ground truth dataset for fact ranking that enables a standardized algorithm comparison and repeatable experimentation. Our contribution also includes the semi-supervised creation of a representative DBpedia sample, the design of statistical formula for calculating ranks of facts, reporting on irregular (nonsense) triples that improve the data quality of DBpedia and the identification of disputed facts which indicate the diversity of opinions. Additionally, we publish the raw data gathered with the tool including sociological and cultural background of users, in order to motivate further research also in related areas (*e. g.* recommender systems, exploratory search). The corpus is constantly growing and in future we will focus on engaging more users, in order to provide a higher quality gold standard with improved heuristics. The data will be versioned on a regular basis.

This work has partially been funded by the German Government, Federal Ministry of Education and Research under the project number 03WKCJ4D.

References

1. B. Bamba and S. Mukherjea. Utilizing resource importance for ranking semantic web query results. *Semantic Web and Databases*, 2005.

[7] `http://s16a.org/kea`

2. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International journal on Semantic Web and Information Systems*, 2009.
3. G. Cheng, T. Tran, and Y. Qu. RELIN: Relatedness and informativeness-based centrality for entity summarization. In *Lecture Notes in Computer Science*, 2011.
4. X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *CoRR*, 2015.
5. T. Franz, A. Schultz, S. Sizov, and S. Staab. TripleRank: Ranking Semantic Web data by tensor decomposition. In *Lecture Notes in Computer Science*, 2009.
6. K. Gunaratna, K. Thirunarayan, and A. P. Sheth. FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
7. J. Hees, T. Roth-Berghofer, R. Biedert, B. Adrian, and A. Dengel. BetterRelations: Collecting association strengths for linked data triples with a game. In *Lecture Notes in Computer Science*. 2012.
8. A. Hogan, A. Harth, and S. Decker. Reconrank: A scalable ranking method for semantic web data with context. In *Second International Workshop on Scalable Semantic Web Knowledge Base Systems, ISWC*, 2006.
9. H. Hwang, V. Hristidis, and Y. Papakonstantinou. ObjectRank: a system for authority-based search on databases. *SIGMOD*, 2006.
10. P. Langer, P. Schulze, S. George, M. Kohnen, T. Metzke, Z. Abedjan, and G. Kasneci. Assigning global relevance scores to DBpedia facts. In *Proceedings of International Conference on Data Engineering*, 2014.
11. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mende, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 2012.
12. H. Paulheim. DBpediaNYD-A Silver Standard Benchmark Dataset for Semantic Relatedness in DBpedia. In *NLP & DBpedia workshop, ISWC*, 2013.
13. M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus Annotation through Crowdsourcing : Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014.
14. N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *IEEE Intelligent Systems*, 2006.
15. A. Singhal. Introducing the Knowledge Graph: things, not strings, 2012.
16. M. Sydow, M. Pikua, and R. Schenkel. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, 2013.
17. A. Thalhammer, M. Knuth, and H. Sack. Evaluating entity summarization using a game-based ground truth. In *Lecture Notes in Computer Science*, 2012.
18. A. Thalhammer and A. Rettinger. Browsing DBpedia Entities with Summaries. In *The Semantic Web: ESWC 2014 Satellite Events*, 2014.
19. A. Thalhammer, I. Toma, A. J. Roa-Valverde, and D. Fensel. Leveraging Usage Data for Linked Data Movie Entity Summarization. *CoRR*, 2012.
20. J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia. *Interactive Technology and Smart Education*, 2011.
21. J. Waitelonis and H. Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, 2012.
22. W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 2010.