GERBIL's New Stunts: Semantic Annotation Benchmarking Improved

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo

AKSW Group, University of Leipzig, Germany
roeder|usbeck|ngonga@informatik.uni-leipzig.de

Abstract. The ability to compare frameworks from the same domain is of central importance for their introduction into complex applications. In the domains of named entity recognition and entity linking, the large number of systems and their orthogonal evaluation w.r.t. measures and datasets has led to an unclear landscape pertaining to the abilities and weaknesses of the different frameworks. With GERBIL—a general framework for benchmarking semantic entity annotation systems—we narrowed this evaluation gap by generating concise, archivable, human- and machine-readable experiments, analytics and diagnostics. In this article, we present how GERBIL addresses the deprecation and inconsistency of existing gold standards for named entity recognition and entity linking, a feature which is currently not supported by the state of the art. We derived the importance of this feature from usage and dataset requirements collected from the GERBIL user community, which has already carried out more than 12,000 single evaluations using our framework. Through the resulting updates, GERBIL now supports 7 tasks, 19 datasets and 13 systems.

1 Introduction

Named Entity Recognition (NER) and Named Entity Linking/Disambiguation (NEL/D) as well as other natural language processing (NLP) tasks play a key role in annotating and retrieving RDF from unstructured data. However, the plethora of NER and NED approaches (see [10] for an overview) as well as their orthogonal evaluation on different datasets using different (definitions of) evaluation measures has resulted in a tool landscape which is difficult to fathom. GERBIL [10] is a generic framework that was designed primarily for benchmarking entity annotation tools with the aim of ensuring repeatable and archiveable experiments following the FAIR principles.[1] It provides (1) an online GUI to configure and run experiments, (2) the assigning of persistent URLs to experiments,[2] and (3) exporting the results of the experiments in human- and machine-readable formats (i.e., RDF). Furthermore, (4) the framework displays the results w.r.t. the data sets and the features of the data sets on which the experiments were performed [9] and (5) offers access to all experiment results via a SPARQL endpoint. GERBIL is an open-source software and available as hosted service.[3]

[1] Findable, Accesible, Interoperatable and Re-Usable, see https://www.force11.org/group/fairgroup/fairprinciples

[2] GERBIL supports W3ID URLs (https://w3id.org/gerbil)

[3] Software: https://github.com/AKSW/gerbil. Service: http://gerbil.aksw.org/gerbil. SPARQL endpoint: http://gerbil.aksw.org/sparql.

After the release of GERBIL and several hundred experiments with 13 datasets, a list of drawbacks of current gold standards stated by GERBIL's community and developers led to requirements for further development of GERBIL. In particular, the requirements pertained to:

- **Entity Matching.** The comparison of two strings representing entity URIs is not sufficient to determine whether an annotator has linked an entity correctly. For example, the two URIs `http://dbpedia.org/resource/Berlin` and `http://en.wikipedia.org/wiki/Berlin` stand for the same real-world object. Hence, the result of an annotation system should be marked as true positive if it generates any of these two URIs to signify the corresponding real-world object. The need to address this drawback of current datasets (which only provide one of these URIs) is amplified by the diversity of annotators and the corresponding diversity of knowledge bases (KB) on which they rely on.
- **Deprecated entities in datasets.** Most of the gold standards in the NER and NED research area have not been updated after their first creation. Thus, the URIs they rely on have remained static over the years while the underlying KBs might have been refined or changed. This leads to some URIs inside a gold standard being deprecated. Like in the first requirement, there is hence a need to provide means to assess a result as true positive when the URI generated by a framework is a novel URI which corresponds to the deprecated URI.
- **New tasks and Adapters** GERBIL has been requested to be used for the two OKE challenges in 2015 and 2016.[4] Thus, we implemented corresponding tasks and supported the execution of the respective campaigns. Additionally, we added 6 state-of-the-art annotators and 6 datasets upon community request.

In the following, we present the developments of GERBIL that address the *URI-based entity matching requirements* of the users. In addition, we present the novel tasks that resulted from the user requirements as well as an evaluation of the evolution of the frameworks contained in GERBIL.

2 Improved URI Matching

The requirements mentioned above are intrinsic to benchmarking entity recognition and linking. Hence, while the solution we present herein is implemented into GERBIL 1.2.2,[5] the approach presented herein is a first attempt towards a general solution to one of the key benchmarking problems underlying NER and NED.

2.1 Entity Matching

The evaluation of entity annotations mainly comprises two parts. An annotation has a position and/or a meaning based on the different experiment types. While (1) the comparison of the positions has already been discussed in former works [10, 3], (2) the

[4] `https://github.com/anuzzolese/oke-challenge` and `https://github.com/anuzzolese/oke-challenge-2016`

[5] `https://github.com/AKSW/Gerbil/releases/tag/v1.2.2`

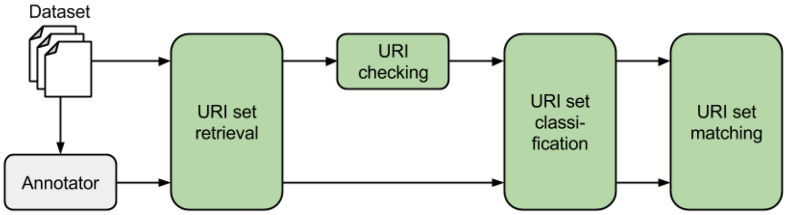

Fig. 1. Schema of the four components of the entity matching process.

evaluation whether two given meanings are matching each other is more challenging. The comparison of two strings representing entity URIs might look like a solution for this problem. However, in practice, this simple approach has various limitations. These limitations are mainly caused by the various ways in which the annotators are expressing their annotation. Some systems are using DBpedia [1] URIs or IRIs while other systems annotate documents with Wikipedia IDs or article titles. Additionally, in most cases the versions of the KBs used to create the datasets are diverging from the versions an annotator relies on.

The key insight behind the solution to this problem in GERBIL is simply to use URIs to represent meanings. We provide an enhanced entity matching which comprise the four steps (1) URI set retrieval, (2) URI checking, (3) URI set classification, and (4) URI set matching, see Figure 1.

URI set retrieval. Since an entity can be described in several KBs using different URIs and IRIs, GERBIL assigns a set of URIs to a single annotation representing the semantic meaning of this annotation. Initially, this set contains the single URI that has been loaded from the dataset or read from an annotators response. The set is expanded by crawling the Semantic Web graph using `owl:sameAs` links as well as redirects. These links are retrieved using different modules that are chosen based on the domain of the URI. A general approach is to de-reference the given URI and try to parse the returned triples. For the DBpedia URIs we offer a module that can transform them into Wikipedia URIs and vice versa. Additionally, we implemented a Wikipedia API client module that can retrieve redirects for Wikipedia URIs. Moreover, one module can handle common errors like wrong domain names, e.g., the usage of `DBpedia.org` instead of `dbpedia.org`, and the transformation from an IRI into a URI and vice versa. The expansion of the set stops, if all URIs in the set have been crawled and no new URI could be added.

URI checking. While the development of annotators moves on, many datasets have been created years ago using versions of KBs that are not used, anymore. This is an important issue that cannot be solved automatically unless the datasets refer to their old versions, which is practically rarely the case. We try to minimize the influence of outdated URIs by checking every URI for its existence. If a URI cannot be dereferenced, it is marked as outdated. However, this is only possible for URIs of KBs that abide by the Linked Data principles and provide de-referencable URIs, e.g., DBpedia.

URI set classification. All entities can be separated into two classes [5]. The class C_{KB} comprises all entities that are present inside at least one KB. In contrast to that, emerging entities are not present in any KB and form the second class C_{EE}. A URI set S is classified as $S \in C_{KB}$ if it contains at least one URI of a predefined KBs namespace. Otherwise it is classified as $S \in C_{EE}$.

URI set matching. The final step of checking whether two entities are matching each other is to check whether their two URI sets are matching. There are two cases in which two URI sets S_1 and S_2 are matching.

$$(S_1 \in C_{KB}) \wedge (S_2 \in C_{KB}) \wedge (S1 \cap S2 \neq \emptyset) \tag{1}$$

$$(S_1 \in C_{EE}) \wedge (S_2 \in C_{EE}) \tag{2}$$

In the first case, both sets are assigned to the C_{KB} class and the sets are overlapping while in the second case, both sets are assigned to the C_{EE} class. Note that in case of emerging entities, it does not make sense to check whether both sets are overlapping since in most cases the URIs of these entities are synthetically generated.

Limitations. This entity matching has two known drawbacks. First, wrong links between KBs can lead to a wrong URI set. The following example shows that because of a wrong linkage between DBpedia and `data.nytimes.com`, Japan and Armenia are the same:[6]

```
 dbr:Japan owl:sameAs nyt:66220885916538669281 .
nyt:66220885916538669281 owl:sameAs dbr:Armenia .
```

Second, the URI set retrieval as well as the URI checking cause a huge communication effort. Since our implementation of this communication is considerate of the KB endpoints by inserting delays between the single requests, these steps slow down the evaluation. However, our future developments will aim at reducing this drawback.

3 Further Developments

3.1 Experiment Types

Since the initial release of GERBIL, we made use of its modular structure and added four tasks to the list of available experiment types. First, we supported the OKE Challenge 2015 [7] by adding the first and second OKE task, i.e, named entity recognition, typing and linking as well as entity type annotation (cf. CETUS [8]), to GERBIL's experiment types. Here, we also integrated a hierarchical F-measure for the evaluation of the new typing task. The challenge requirements led to the introduction of the new concept of sub-tasks. Thus, users are now able to analyze which steps of their pipeline leads to

[6] `dbr` is the prefix for `http://dbpedia.org/resource/`, `owl` is the prefix for `http://www.w3.org/2002/07/owl#` and `nyt` is the prefix for `http://data.nytimes.com/`.

errors, e.g., whether the linking or the recognition step of an annotator caused the most problems. Second, we directly derived two tasks, namely Entity Recognition and Entity Typing from the two tasks above.

We removed the separation between Sa2KB and A2KB as well as between Sc2KB and C2KB. In the new version, the usage of confidence scores is a part of the A2KB and C2KB experiments. If an annotator adds confidence scores to its annotations, GERBIL searches a threshold that optimizes the micro F1-score.

3.2 Improved Diagnostics

To support the development of new approaches, we implemented additional diagnostic capabilities such as the calculation of correlations of dataset features and annotator performance [9]. Figure 2 shows the correlations which can help to figure out strengths and weaknesses of the different approaches.

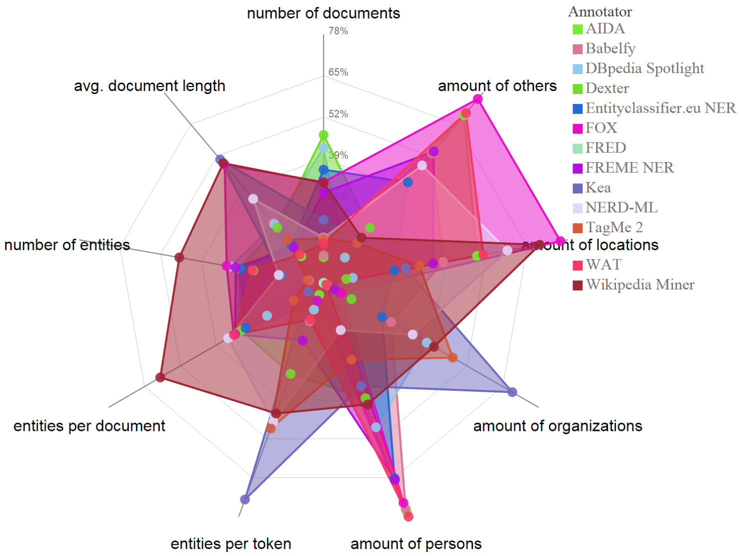

Fig. 2. Absolute correlation values of the annotators Micro F1-scores and the dataset features for the A2KB experiment and the weak annotation match (Date: 13.04.2016).

For a more detailed analysis of the annotator performance, we implemented the possibility to add new metrics to the evaluation, e.g., runtime measurements. Moreover, we added different performance measures that focus on specific parts of the tasks. Beside the general Micro and Macro Precision, Recall and F1-measure, GERBIL offers three other measures that take the classification of the entities into account. Table 1 shows the different cases that can occur when sets of URIs are compared.

While all cases are taken into account for the normal measures, the *InKB* measures focus on those cases in which either the URI set of the dataset or the URI set of the

Table 1. The different classification cases that can occur during the evaluation. A dash means that there is no URI set that could be used for the matching. A tick shows that this case is taken into account while calculating the measure.

Dataset	Annotator	Normal	InKB	EE	GSInKB
$S_1 \in KB$	$S_2 \in KB$	✓	✓		✓
$S_1 \in KB$	$S_2 \notin KB$	✓	✓	✓	✓
$S_1 \in KB$	—	✓	✓		✓
$S_1 \notin KB$	$S_2 \in KB$	✓	✓	✓	
$S_1 \notin KB$	$S_2 \notin KB$	✓		✓	
$S_1 \notin KB$	—	✓		✓	
—	$S_2 \in KB$	✓	✓		
—	$S_2 \notin KB$	✓		✓	

annotator are classified as $S \in C_{KB}$. The same holds for the *EE* measures and the C_{EE} class. Both measures can be used to check the performance for one of these two classes. The *GSInKB* measures are only calculated for NED experiments (D2KB). It can be used to assess the performance of an annotator if there where no emerging entities inside the dataset.

3.3 New Datasets and Annotators

The implementation of the OKE challenge tasks also added six new datasets designed for the challenge.The datasets were manually created and approved by at least two domain experts and contain NIF-based annotations for RDF entities and classes, cf. the OKE challenge documentation for further details [7]. Overall, GERBIL now contains 19 individual datasets available to 7 experiment types.

The current version 1.2.2 contains six new annotators. While CETUS, CETUS_FOX [8] and FRED [2] were added to take part in the OKE challenge 2015, AIDA [6], FREME e-Entity[7] and `entityclassifier.eu` [4] were added to enhance the spectrum of A2KB annotators.

4 Contribution from and to the Community

One of GERBILs main goals was to provide the community with an online benchmarking tool that provides archivable and comparable experiment URIs. Thus, the impact of the framework can be measured by analyzing the interactions on the platform itself. Since its first public release on the 17th October 2014 until the 15th February 2015, 1.824 experiments were started on the platform containing more than 12.466 tasks for annotator-dataset pairs. According to our mail correspondence, we can deduce that more than 20 local installations of GERBIL exist for testing novel annotation systems both in industry and academia. This shows the intensive usage of our GERBIL instance. One interesting aspect is the usage of the different provided systems, especially the

[7] `https://github.com/freme-project/e-Entity`

Table 2. Number of tasks executed per annotator. By caching results we did not need to execute 12466 tasks but only 9906.

Annotator	Number of Tasks
NIF-based Annotators	2519
Babelfy	958
DBpedia Spotlight	922
TagMe 2	811
WAT	787
Kea	763
Wikipedia Miner	714
NERD-ML	639
Dexter	587
AGDISTIS	443
Entityclassifier.eu NER	410
FOX	352
Cetus	1

heavy exploitation of the possibility to test NIF-based webservices, see Table 2. Thus, GERBIL is a powerful evaluation tool for developing new annotators that can be evaluated easily by using the NIF-based interface. Moreover, another impact of the GERBIL platform can be seen directly by observing annotator performance over time.

By archiving experiments, we are able to measure the time-dependent evolution of systems for the first time. Thus, we can clarify research questions such as 1) Is there an active development or 2) Does a system update influence evaluations? Figure 3 shows an A2KB experiment over more than a year on the MSNBC dataset measuring 10 online available annotators. The observations show that a) some systems are unreachable and then recover quickly (e.g., beginning of 2015). Thus, we assume that there is an active community maintaining the projects. b) Systems do improve over time (such as Wikipedia Miner, DBpedia Spotlight) and that c) our system updates can influence a systems performance due to changes in the underlying measurement infrastructure (autumn 2015).

5 Conclusion & Future work

We presented the novel version of GERBIL—an improved platform for repeatable, storable and citable semantic annotation experiments—and how we extended it since its release. We pointed to the general problem of the matching of entities and presented our current solution. GERBIL has been adapted by the community and used for many experiments and publications. Especially, the broad usage of the NIF-based webservice interface shows that GERBIL can be used to evaluate prototypes and, thus, enhances and speeds up the development of new systems. In future works, GERBIL will be used as a basis for a holistic benchmarking platform for (Big) Linked Data and thus be extended towards novel, industry-driven measures, datasets and annotators. Moreover, GERBIL will support the benchmarking efforts of future QALD challenges.

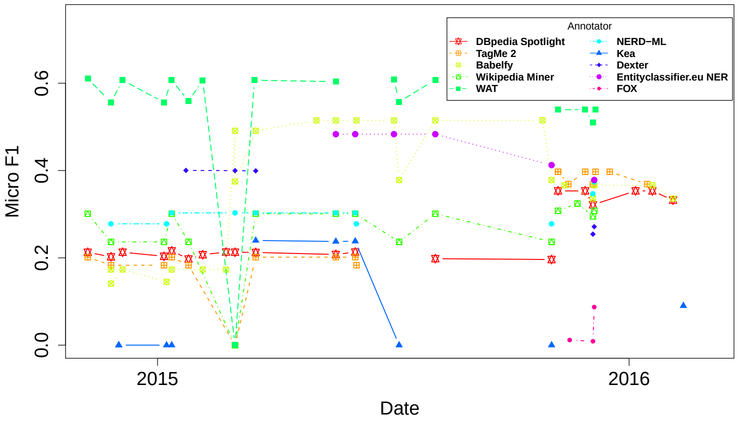

Fig. 3. Different annotators executed on the small dataset MSNBC from the first deployment to the 4th February 2016.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the ISWC 2008*. Springer, 2008.
2. S. Consoli and D. Recupero. Using FRED for Named Entity Resolution, Linking and Typing for Knowledge Base Population. In *Semantic Web Evaluation Challenges*, volume 548 of *Communications in Computer and Information Science*, pages 40–50. 2015.
3. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 249–260. IW3C2, 2013.
4. M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-time classification of entities in text with wikipedia. In *Proceedings of the ECMLPKDD'13*, 2013.
5. J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 385–396, New York, NY, USA, 2014. ACM.
6. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, pages 782–792, 2011.
7. A.-G. Nuzzolese, A. Gentile, V. Presutti, A. Gangemi, D. Garigliotti, and R. Navigli. Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*, volume 548 of *Communications in Computer and Information Science*, pages 3–15. Springer International Publishing, 2015.
8. M. Röder, R. Usbeck, R. Speck, and A.-C. Ngonga Ngomo. CETUS – A Baseline Approach to Type Extraction. In *1st Open Knowledge Extraction Challenge at 12th European Semantic Web Conference (ESWC 2015)*, 2015.
9. R. Usbeck, M. Röder, and A.-C. Ngonga Ngomo. Evaluating Entity Annotators Using GERBIL. In *Proceedings of 12th European Semantic Web Conference (ESWC 2015)*, 2015.
10. R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*, 2015.