Semantic Web 0 (2016) 1
IOS Press

GERBIL – Benchmarking Named Entity Recognition and Linking Consistently

Michael Röder *, Ricardo Usbeck and Axel-Cyrille Ngonga Ngomo
AKSW, Leipzig University, Germany
E-mail: {usbeck,roeder,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. We present GERBIL—an improved platform for repeatable, storable and citable semantic annotation experiments—and how we extended it since its release. With GERBIL, we narrowed this evaluation gap by generating concise, archivable, human- and machine-readable experiments, analytics and diagnostics. The rationale behind our framework is to provide developers, end users and researchers with easy-to-use interfaces that allow for the agile, fine-grained and uniform evaluation of annotation tools on multiple datasets. By these means, we aim to ensure that both tool developers and end users can derive meaningful insights pertaining to the extension, integration and use of annotation applications. In particular, GERBIL provides comparable results to tool developers so as to allow them to easily discover the strengths and weaknesses of their implementations with respect to the state of the art. With the permanent experiment URIs provided by our framework, we ensure the reproducibility and archiving of evaluation results. Moreover, the framework generates data in machine-processable format, allowing for the efficient querying and post-processing of evaluation results. Additionally, the tool diagnostics provided by GERBIL allows deriving insights pertaining to the areas in which tools should be further refined, thus allowing developers to create an informed agenda for extensions and end users to detect the right tools for their purposes. Finally, we implemented additional types of experiments including entity typing. GERBIL aims to become a focal point for the state of the art, driving the research agenda of the community by presenting comparable objective evaluation results. Furthermore, we tackle the central problem of the evaluation of entity linking, i.e., we answer the question how an evaluation algorithm can compare two URIs to each other without being bound to a specific knowledge base. Our approach to this problem opens a way to address the deprecation of URIs 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 24.000 single evaluations using our framework. Through the resulting updates, GERBIL now supports 8 tasks, 46 datasets and 20 systems.

Keywords: Semantic Entity Annotation System, Reusability, Archivability, Benchmarking Framework, Named Entity Recognition, Linking, Disambiguation

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 RDF knowledge from unstructured data. While manifold annotation tools have been developed over the last years to address (some of) the subtasks related to the extraction of structured data from unstructured data [18,26,36,39,41,47,51,58,61], the provision of comparable results for these tools remains a tedious problem. The issue of comparability of results is not to be regarded as being intrinsic to the annotation task. Indeed, it is now well established that scientists spend between 60 and 80% of their time

*Corresponding author. E-mail: roeder@informatik.uni-leipzig.de

1570-0844/16/$35.00 © 2016 – IOS Press and the authors. All rights reserved

preparing data for experiments [21,28,46]. Data preparation being such a tedious problem in the annotation domain is mostly due to the different formats of the gold standards as well as the different data representations across reference datasets. These restrictions have led to authors evaluating their approaches on datasets (1) that are available to them and (2) for which writing a parser as well as of an evaluation tool can be carried out with reasonable effort. In addition, a large number of quality measures have been developed and used actively across the annotation research community to evaluate the same task, leading to the results across publications on the same topics not being easily comparable. For example, while some authors publish macro-F-measures and simply call them F-measures, others publish micro-F-measures for the same purpose, leading to significant discrepancies across the scores. The same holds for the evaluation of how well entities match. Indeed, partial matches and complete matches have been used in previous evaluations of annotation tools [9,56]. This heterogeneous landscape of tools, datasets and measures leads to a poor repeatability of experiments, which makes the evaluation of the real performance of novel approaches against the state of the art rather difficult.

Thus, we present GERBIL—a general framework for benchmarking semantic entity annotation systems which introduces a platform and a software for comparable, archivable and efficient semantic annotation experiments fostering a more efficient and effective community.[1]

In the rest of this paper, we explain the core principles which we followed to create GERBIL and detail our new contributions. Thereafter, we present the state of the art in benchmarking Named Entity Recognition, Typing and Linking. In Section 4, we present the GERBIL framework. We focus in particular on the provided features such as annotators, datasets, metrices and the evaluation processes including our new approach to match URIs. We then present an evaluation of the framework by indirectly qualifying the interaction of the community with our platform since its release. We conclude with a discussion of the current state of GERBIL and a presentation of future work. More information can be found at our project webpage `http://gerbil.aksw.org` and at the code repository page `https://github.com/AKSW/gerbil`. The online version of GERBIL can be accessed at `http://gerbil.aksw.org/gerbil`.

[1] This paper is a significant extension of [62] including the progress of the GERBIL project since its initial release in 2015.

2. Principles

The insights on the difficulties of current evaluation setups have led to a movement towards the creation of frameworks to ease the evaluation of solutions that address the same annotation problem, see Section 3. GERBIL is a community-driven effort to enable the continuous evaluation of annotation tools. Our approach is an open-source, extensible framework that allows evaluating tools against (currently) 20 different annotators on 12 different datasets within 6 different experiment types. By integrating such a large number of datasets, experiment types and frameworks, GERBIL allows users to evaluate their tools against other semantic entity annotation systems (short: entity annotation systems) by using exactly the same setting, leading to fair comparisons based on exactly the same measures. Our approach goes beyond the state of the art in several respects:

- **Repeatable settings**: GERBIL provides *persistent URLs* for experimental settings. Hence, by using GERBIL for experiments, tool developers can ensure that the settings for their experiments (measures, datasets, versions of the reference frameworks, etc.) can be reconstructed in a unique manner in future works.
- **Archivable experiments**: Through experiment URLs, GERBIL also addresses the problem of archiving experimental results and allows end users to gather all pieces of information required to choose annotation frameworks for practical applications.
- **Open software and service**: GERBIL aims to be a *central repository for annotation results* without being a central point of failure: While we make experiment URLs available, we also provide users directly with their results to ensure that they use them locally without having to rely on GERBIL.
- **Leveraging RDF for storage**: The results of GERBIL are published in a *machine-readable format*. In particular, our use of DataID [3] and DataCube [12] to denote tools and datasets ensures that results can be easily combined and queried (for example to study the evolution of the performance of frameworks) while the exact configuration of the experiments remains uniquely re-

constructable. By these means, we also tackle the problem of *reproducibility*.
- **Fast configuration**: Through the provision of results on different datasets of different types and the provision of results on a simple user interface, GERBIL also provides means to quickly gain an overview of the current performance of annotation tools, thus providing (1) developers with insights pertaining to the type of data on which their accuracy needs improvement and (2) end users with insights allowing them to choose the right tool for the tasks at hand.
- **Any knowledge base**: With GERBIL we introduce the notion of knowledge base-agnostic benchmarking of entity annotation systems through generalized experiment types. By these means, we allow benchmarking tools against reference datasets from any domain grounded in any reference knowledge base.

To ensure that the GERBIL framework is useful to both end users and tool developers, its architecture and interface were designed with the following requirements in mind:

- **Easy integration of annotators**: We provide a wrapping interface that allows annotators to be evaluated via their REST interface. In particular, we integrated 15 additional annotators not evaluated against each other in previous works (e.g., [9]).
- **Easy integration of datasets**: We also provide means to gather datasets for evaluation directly from data services such as DataHub.[2] In particular, we added 37 new datasets to GERBIL.
- **Easy addition of new measures**: The evaluation measures used by GERBIL are implemented as interfaces. Thus, the framework can be easily extended with novel measures devised by the annotation community.
- **Extensibility**: GERBIL is provided as an open-source platform[3] that can be extended by members of the community both to new tasks and different purposes.
- **Diagnostics**: The interface of the tool was designed to provide developers with means to easily detect aspects in which their tool(s) need(s) to be improved.
- **Portability of results**: We generate human- and machine-readable results to ensure maximum usefulness and portability of the results generated by our framework.

After the release of GERBIL and several hundred experiments, a list of drawbacks of current datasets stated by GERBIL's community and developers led to requirements for further development of the platform. 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 several state-of-the-art annotators and datasets upon community request.

Finally, GERBIL was designed primarily for benchmarking entity annotation tools with the aim of ensuring repeatable and archiveable experiments following

[2] `http://datahub.io`
[3] Available at `http://gerbil.aksw.org`.
[4] `https://github.com/anuzzolese/oke-challenge` and `https://github.com/anuzzolese/oke-challenge-2016`

the FAIR principles [65]. Table 1 depicts the details of GERBIL's implementation of the FAIR principles.

3. Related Work

Named Entity Recognition and Entity Linking have gained significant momentum with the growth of Linked Data and structured knowledge bases. Over the last few years, the problem of result comparability has thus led to the development of a handful of frameworks.

The BAT-framework [9] is designed to facilitate the benchmarking of NER, NEL/D and concept tagging approaches. BAT compares seven existing entity annotation approaches using Wikipedia as reference. Moreover, it defines six different task types, five different matchings and six evaluation measures providing five datasets. Rizzo et al. [51] present a state-of-the-art study of NER and NEL systems for annotating newswire and micropost documents using well-known benchmark datasets, namely CoNLL2003 and Microposts 2013 for NER as well as AIDA/CoNLL and Microposts2014 [4] for NED. The authors propose a common schema, named the NERD ontology,[5] to align the different taxonomies used by various extractors. To tackle the disambiguation ambiguity, they propose a method to identify the closest DBpedia resource by (exact-)matching the entity mention. Recently, Chen et al. [27] published EUEF, the easy-to-use evaluation framework which addresses three more challenges as opposed to the standard GERBIL algorithm. First, EUEF introduces a new matching metric based on fuzzy matching to account for annotator mistakes. Second, the framework introduces a new methodology for handling NIL annotations. Third, Chen et al.'s framework analyzes also sub-components of NER/NED systems. However, EUEF only includes three systems and seven datasets and is not open source or online available yet.

Over the course of the last 25 years several challenges, workshops and conferences dedicated themselves to the comparable evaluation of information extraction (IE) systems. Starting in 1993, the Message Understanding Conference (MUC) introduced a first systematic comparison of information extraction approaches [59]. Ten years later, the Conference on Computational Natural Language Learning (CoNLL) started to offer a shared task on named entity recognition and published the CoNLL corpus [60]. In addition, the Automatic Content Extraction (ACE) challenge [14], organized by NIST, evaluated several approaches but was discontinued in 2008. Since 2009, the text analytics conference hosts the workshop on knowledge base population (TAC-KBP) [34] where mainly linguistic-based approaches are published. The Senseval challenge, originally concerned with classical NLP disciplines, has widened its focus in 2007 and changed its name to SemEval to account for the recently recognized impact of semantic technologies [29]. The Making Sense of Microposts workshop series (#Microposts) established in 2013 an entity recognition and in 2014 an entity linking challenge thereby focusing on tweets and microposts [54]. In 2014, Carmel et al. [6] introduced one of the first Web-based evaluation systems for NER and NED and the centerpiece of the entity recognition and disambiguation (ERD) challenge. Here, all frameworks are evaluated against the same unseen dataset and provided with corresponding results.

GERBIL goes beyond the state of the art by extending the BAT-framework as well as [51] in several dimensions to enhance reproducibility, diagnostics and publishability of entity annotation systems. In particular, we provide 37 additional datasets and 15 additional annotators. The framework addresses the lack of treatment of NIL values within the BAT-framework and provides more wrapping approaches for annotators and datasets. Moreover, GERBIL provides persistent URLs for experiment results, unique URIs for frameworks and datasets, a machine-readable output and automatic dataset updates from data portals. Thus, it allows for a holistic comparison of existing annotators while simplifying the archiving of experimental results. Moreover, our framework offers opportunities for the fast and simple evaluation of entity annotation system prototypes via novel NIF-based [23] interfaces, which are designed to simplify the exchange of data and binding of services.

4. The GERBIL Framework

4.1. Architecture Overview

GERBIL abides by a service-oriented architecture driven by the model-view-controller pattern (see Figure 1). Entity annotation systems, datasets and configurations like experiment type, matching or measure

[5] http://nerd.eurecom.fr/ontology

Table 1
FAIR principles and how GERBIL addresses each of them.

To be Findable	
F1. (meta)data are assigned a globally unique and persistent identifier	Unique W3ID URIs per experiment
F2. data are described with rich metadata (defined by R1 below)	Experimental configuration as RDF
F3. metadata clearly and explicitly include the identifier of the data it describes	Relates via RDF
F4. (meta)data are registered or indexed in a searchable resource	batch-updated SPARQL endpoint: `http://gerbil.aksw.org/sparql`
To be Accessible	
A1. (meta)data are retrievable by their identifier using a standardized communications protocol	HTTP (with JSON-LD as data format)
A1.1 the protocol is open, free, and universally implementable	HTTP is an open standard
A1.2 the protocol allows for an authentication and authorization procedure, where necessary	Not necessary, see GERBIL disclaimer
A2. metadata are accessible, even when the data are no longer available	Each experiment is archived
To be Interoperable	
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	RDF, DataID, DataCube
I2. (meta)data use vocabularies that follow FAIR principles	Community-based, open vocabularies
I3. (meta)data include qualified references to other (meta)data	Datasets are described using DataID
To be Reusable	
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	Experiment measures have been chosen in a community process
R1.1. (meta)data are released with a clear and accessible data usage license	GERBIL is implement by LGPL-3.0
R1.2. (meta)data are associated with detailed provenance	Provenance is added to each machine-readable experiment data
R1.3. (meta)data meet domain-relevant community standards	GERBIL covers a superset of domain-relevant data

are implemented as controller interfaces easily pluggable to the core controller. The output of experiments as well as descriptions of the various components are stored in a serverless database for fast deployment. Finally, the view component displays configuration options respectively renders experiment results delivered by the main controller communication with the diverse interfaces and the database.

4.2. Features

Experiments run in our framework can be configured in several manners. In the following, we present some of the most important parameters of experiments available in GERBIL.

4.2.1. Experiment types

An experiment type defines the way used to solve a certain problem when extracting information. Cornolti et al.'s [9] BAT-framework offers six different experiment types, namely (scored) annotation (S/A2KB), disambiguation (D2KB)—also known as linking—and (scored respectively ranked) concept annotation (S/R/C2KB) of texts. In [51], the authors propose two types of experiments, focusing on highlighting the strengths and weaknesses of the analyzed systems. Thereby, performing *i)* entity recognition, i.e., the detection of the exact match of the pair entity mention and type (e.g., detecting the mention *Barack Obama* and typing it as a *Person*), and *ii)* entity linking, where an exact match of the mention is given and the associated DBpedia URI has to be linked (e.g., locating a resource in DBpedia which describes the mention *Barack Obama*). This work differs from the previous one for experimenting in entity recognition, and on annotating entities to a RDF knowledge base.

GERBIL merges the six experiments provided by the BAT-framework to three experiment types by a general handling of scored annotations. These experiment

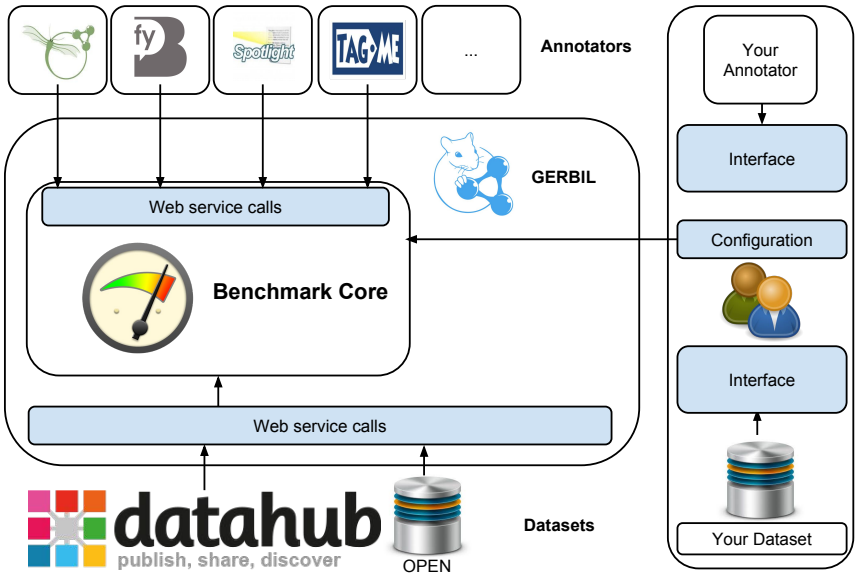

Fig. 1. Overview of GERBIL's abstract architecture. Interfaces to users and providers of datasets and annotators are marked in blue.

types are further extended by the idea to not only link to Wikipedia but to any knowledge base K. One major formal update of the measures in GERBIL is that in addition to implementing experiment types from previous frameworks, it also measures the influence of emerging entities (EEs or NIL annotations), i.e., the linking of entities that are recognized as such but cannot be linked to any resource from the reference knowledge base K. For example, the string *"Ricardo Usbeck"* can be recognized as a person name by several tools but cannot be linked to Wikipedia/DBpedia, as it does not have a URI in these reference datasets. Our framework extends the experiments types of [9] as follows: Let $m = (s, l, d, c) \in M$ denote an entity mention in document $d \in D$ with start position s, length l and confidence score $c \in [0, 1]$. Note that some frameworks might not return (1) a position s or a length l for a mention, in which case we set $s = 0$ and $l = 0$; (2) a score c, in which case we set $c = 1$.

We implement 8 types of experiments:

1. **Entity Recognition**: In this task the entity mentions need to be extracted from a document set D. To this end, an extraction function $ex : D \rightarrow 2^M$ must be computed.
2. **D2KB**: The goal of this experiment type is to map a set of *given* entities mentions (i.e., a subset $\mu \subseteq M$) to entities from a given knowledge base or to NIL. Formally, this is equivalent to finding a mapping $a : \mu \rightarrow K \cup \{NIL\}$. In the classical setting for this task, the start position, the length and the score of the mentions m_i are not taken into consideration.
3. **Entity Typing**: The typing task is similar to the D2KB task. Its goal is to map a set of *given* entities mentions μ to the type hierarchy of K. This task uses the hierarchical F-measure to evaluate the types returned by the annotation system using the expected types of the gold standard and the type hierarchy of K.
4. **C2KB**: The concept tagging task C2KB aims to detect entities when given a document. Formally, the tagging function *tag* simply returns a subset of K for each input document d.
5. **A2KB**: This task is the classical NER/D task, thus a combination of the Entity Recognition and D2KB tasks. Thus, an A2KB annotation system receives the document set D, has to identify entities mentions μ and link them to K.
6. **RT2KB**: This task is the combincation of the Entity Recognition and Typing tasks, i.e., the goal is to identify entities in a given document set D and map them to the types of K.
7. **OKE 2015 Task1**: The first task of the OKE Challenge 2015 [45] comprises the tasks Entity Recognition, Entity Typing and D2KB.
8. **OKE 2015 Task2**: The goal of the second task of the OKE Challenge 2015 [45] is to extract the part of the text that contains the type of a given

entity mention and link it to the type hierarchy of K.

With this extension, our framework can now deal with gold standard datasets and annotators that link to any knowledge base, e.g., DBpedia, BabelNet [44] etc., as long as the necessary identifiers are URIs. We were thus able to implement 37 new gold standard datasets, cf. Section 4.4, and 15 new annotators linking entities to any knowledge base instead of solely Wikipedia like in previous works, cf. Section 4.3.1. With this extensible interface, GERBIL can be extended to deal with supplementary experiment types, e.g., entity salience [9], word sense disambiguation (WSD) [41] and relation extraction [56]. These categories of experiment types will be added to GERBIL in next versions.

4.2.2. Matching

A matching defines which conditions the result of an annotator has to fulfill to be a correct result, i.e., to match an annotation of the gold standard. An annotation has either a position, a meaning (i.e., a linked entity or a type) or both. Therefore, we can define an annotation $a = (s, l, d, u)$ with a start position s and a length l as defined above. d is the document the annotation belongs to and u is a URI that is the link to an entity or the type of an entity (depending on the experiment type).

The first matching type M_e used for the C2KB experiments is the *strong entity matching*. This matching does not rely on positions and takes only the URIs u into account. Following this matching, a single annotation $a = (s, l, d, u)$ returned by the annotator is correct iff it matches exactly with one of the annotations $a' = (s', l', d, u')$ in the gold standard $a' \in G(d)$ of d [9]. Formally,

$$M_e(a, a') = \begin{cases} 1 & \text{iff } u = u', \\ 0 & \text{else.} \end{cases} \tag{1}$$

For the D2KB experiments, the matching is expanded to the *strong annotation matching* M_a and includes the correct position of the entity mention inside the document:

$$M_a(m, G) = \begin{cases} 1 & \text{iff } u = u' \wedge s = s' \\ & \wedge l = l', \\ 0 & \text{else.} \end{cases} \tag{2}$$

The strong annotation matching can be used for A2KB and Sa2KB experiments, too. However, in practice this exact matching can be misleading. A document can contain a gold standard named entity like *"President Barack Obama"* while the result of an annotator only marks *"Barack Obama"* as named entity. Using an exact matching leads to weighting this result as wrong while a human might rate it as correct. Therefore, the *weak annotation matching* M_w relaxes the conditions of the strong annotation matching. Thus, a correct annotation has to be linked to the same entity and must overlap the annotation of the gold standard:

$$M_w(m, G) = \begin{cases} 1 & \text{iff } u = u' \\ & \wedge((s \leq s' \wedge e \leq e') \\ & \vee(s \geq s' \wedge e \geq e') \\ & \vee(s \leq s' \wedge e \geq e') \\ & \vee(s \geq s' \wedge e \leq e')) \\ 0 & \text{else} \end{cases} \tag{3}$$

where $e = s + l$ and $e' = s' + l'$.

However, the evaluation whether two given meanings are matching each other is more challenging than the expression $u = u'$ reveals. The comparison of two strings representing entity URIs might look like a solution for this problem. However, in practice, this simple approach has limitations. These limitations are mainly caused by the various ways in which the annotators are expressing their annotation. Some systems are using DBpedia [31] 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 2.

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. The general ap-

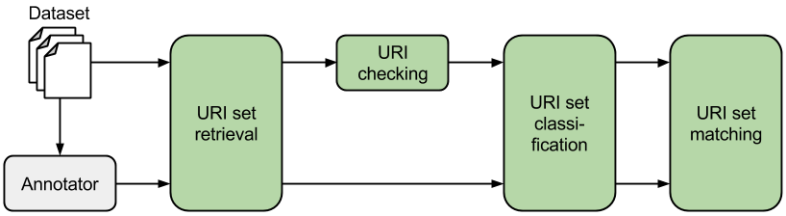

Fig. 2. Schema of the four components of the entity matching process.

proach we implemented de-references the given URI and tries to parse the returned triples. Although this approach works with every KB, we offer a module for the DBpedia URIs 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 [24]. 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.

$$(S_1 \in C_{KB}) \wedge (S_2 \in C_{KB}) \wedge (S1 \cap S2 \neq \emptyset) \quad (4)$$

$$(S_1 \in C_{EE}) \wedge (S_2 \in C_{EE}) \quad (5)$$

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.

4.2.3. Measures

GERBIL comes with six measures subdivided into two groups and derived from the BAT-framework, namely the micro- and the macro-group of precision, recall and f-measure. 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 mea-

[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/`.

Table 2

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
$S_1 \in KB$	$S_2 \in KB$	✓	✓		✓
$S_1 \in KB$	$S_2 \notin KB$	✓	✓	✓	✓
$S_1 \in KB$	—	✓	✓		✓
$S_1 \notin KB$	$S_2 \in KB$	✓	✓	✓	
$S_1 \notin KB$	$S_2 \notin KB$	✓		✓	
$S_1 \notin KB$	—	✓		✓	
—	$S_2 \in KB$	✓	✓		
—	$S_2 \notin KB$	✓		✓	

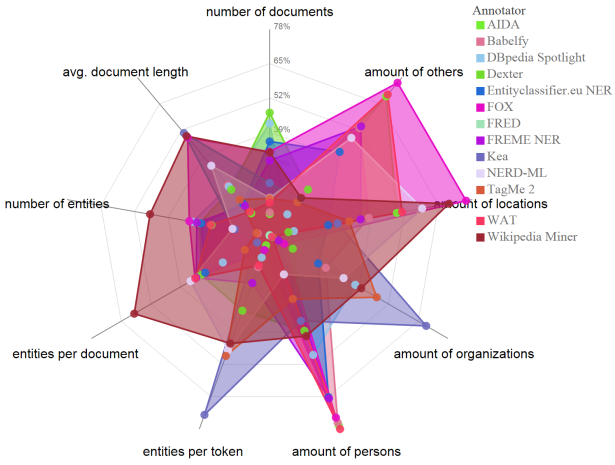

Fig. 3. Absolute correlation values of the annotators Micro F1-scores and the dataset features for the A2KB experiment and the weak annotation match (Date: 13.04.2016).

sures that take the classification of the entities into account. Table 2 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 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 as if there where no emerging entities inside the dataset, e.g., if the annotation system is not capable of handling these entities.

4.3. 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 [63]. Figure 3 shows the correlations which can help to figure out strengths and weaknesses of the different approaches.

4.3.1. Annotators

GERBIL aims to reduce the amount of work required to compare existing as well as novel annotators in a comprehensive and reproducible way. To this end, we provide two main approaches to evaluating entity annotation systems with GERBIL.

1. **BAT-framework Adapter**
   Within BAT, annotators can be implemented by wrapping using a Java-based interface. GERBIL offers an adapter so the wrappers of the BAT-framework can be reused easily. Due to the community effort behind GERBIL, we could raise the number of published annotators from 5 to 20. We investigated the effort to implement a BAT-framework adapter in contrast to evaluation efforts done without a structured evaluation framework in Section 5.
2. **NIF-based Services**: GERBIL implements means to understand NIF-based [23] communication over web-service in two ways. First, if the server-side implementation of annotators understands NIF-documents as input and output format, GERBIL and the framework can simply exchange NIF-documents.[7] Thus, novel NIF-based annotators can be deployed efficiently into GERBIL and use a more robust communication format compared to the amount of work necessary for deploying and writing a BAT-framework adapter. Second, if developers do not want to publish their APIs or write source code, GERBIL offers the possibility for NIF-based webservices to be tested online by providing their URI and name only.[8] GERBIL does not store these connections in terms of API keys or URLs but still offers the opportunity of persistent experiment results.

[7] We describe the exact requirements to the structure of the NIF document on our project website's wiki as NIF offers several ways to build a NIF-based document or corpus.

[8] `http://gerbil.aksw.org/gerbil/config`

Currently, GERBIL offers 20 entity annotation systems with a variety of features, capabilities and experiments. In the following, we present current state-of-the-art approaches both available or unavailable in GERBIL.

1. **Cucerzan**: As early as in 2007, Cucerzan presented a NED approach based on Wikipedia [11]. The approach tries to maximize the agreement between contextual information of input text and a Wikipedia page as well as category tags on the Wikipedia pages. The test data is still available[9] but since we can safely assume that the Wikipedia page content changed a lot since 2006, we do not use it in our framework, nor we are aware of any publication reusing this data. Furthermore, we were not able to find a running webservice or source code for this approach.
2. **Wikipedia Miner**: This approach was introduced in [39] in 2008 and is based on different facts like prior probabilities, context relatedness and quality, which are then combined and tuned using a classifier. The authors evaluated their approach based on a subset of the AQUAINT dataset.[10] They provide the source code for their approach as well as a webservice[11] which is available in GERBIL.
3. **Illinois Wikifier**: In 2011, [49] presented an NED approach for entities from Wikipedia. In this article, the authors compare local approaches, e.g., using string similarity, with global approaches, which use context information and lead finally to better results. The authors provide their datasets[12] as well as their software "Illinois Wikifier"[13] online. Since "Illinois Wikifier" is currently only available as local binary and GERBIL is solely based on webservices we excluded it from GERBIL for the sake of comparability and server load.
4. **DBpedia Spotlight**: One of the first semantic approaches [36] was published in 2011, this framework combines NER and NED approach based upon DBpedia.[14] Based on a vector-space representation of entities and using the cosine similarity, this approach has a public (NIF-based) webservice[15] as well as its online available evaluation dataset.[16]
5. **AIDA**: The AIDA approach [26] relies on coherence graph building and dense subgraph algorithms and is based on the YAGO2[17] knowledge base. Although the authors provide their source code, a webservice and their dataset which is a manually annotated subset of the 2003 CoNLL share task [60], GERBIL will not use the webservice since it is not stable enough for regular replication purposes at the moment of this publication.[18] That is, the AIDA team discourages the use because they constantly switch the underlying entity repository, and tune parameters.
6. **TagMe 2**: TagMe 2 [18] was published in 2012 and is based on a directory of links, pages and an inlink graph from Wikipedia. The approach recognizes named entities by matching terms with Wikipedia link texts and disambiguates the match using the in-link graph and the page dataset. Afterwards, TagMe 2 prunes the identified named entities which are considered as non-coherent to the rest of the named entities in the input text. The authors publish a key-protected webservice[19] as well as their datasets[20] online. The source code, licensed under Apache 2 licence can be obtained directly from the authors. The datasets comprise only fragments of 30 words and less of full documents and will not be part of the current version of GERBIL.
7. **NERD-ML**: In 2013, [17] proposed an approach for entity recognition tailored for extracting entities from tweets. The approach relies on a machine learning classification of the entity type given a rich feature vector composed of a set of linguistic features, the output of a properly

[9] http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/
[10] http://www.nist.gov/tac/data/data_desc.html#AQUAINT
[11] http://wikipedia-miner.cms.waikato.ac.nz/
[12] http://cogcomp.cs.illinois.edu/page/resource_view/4
[13] http://cogcomp.cs.illinois.edu/page/software_view/33
[14] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Known-uses
[15] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service
[16] http://wiki.dbpedia.org/spotlight/isemantics2011/evaluation
[17] http://www.mpi-inf.mpg.de/yago-naga/yago/
[18] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/
[19] http://tagme.di.unipi.it/
[20] http://acube.di.unipi.it/tagme-dataset/

trained Conditional Random Fields classifier and the output of a set of off-the-shelf NER extractors supported by the NERD Framework. The follow-up, NERD-ML [51], improved the classification task by re-designing the selection of the features. The authors assessed the NERD-ML's performance on both microposts and newswire domains. NERD-ML has a public webservice which is part of GERBIL.[21]

8. **KEA NER/NED**: This approach is the successor of the approach introduced in [58] which is based on a fine-granular context model taking into account heterogeneous text sources as well as text created by automated multimedia analysis. The source texts can have different levels of accuracy, completeness, granularity and reliability which influence the determination of the current context. Ambiguity is solved by selecting entity candidates with the highest level of probability according to the predetermined context. The new implementation begins with the detection of groups of consecutive words (n-gram analysis) and a lookup of all potential DBpedia candidate entities for each n-gram. The disambiguation of candidate entities is based on a scoring cascade. KEA is available as NIF-based webservice.[22]
9. **WAT**: WAT is the successor of TagME [18].[23] The new annotator includes a re-design of all TagME components, namely, the spotter, the disambiguator, and the pruner. Two disambiguation families were newly introduced: graph-based algorithms for collective entity linking based and vote-based algorithms for local entity disambiguation (based on the work of Ferragina et al. [18]). The spotter and the pruner can be tuned using SVM linear models. Additionally, the library can be used as a D2KB-only system by feeding appropriate mention spans to the system.
10. **Dexter**: This approach [7] is an open-source implementation of an entity disambiguation framework. The system was implemented in order to simplify the implementation of an entity linking approach and allows to replace single parts of the process. The authors implemented several state-of-the-art disambiguation methods. Results in this paper are obtained using an implementation of the original TagMe disambiguation function. Moreover, Ceccarelli et al. provide the source code[24] as well as a webservice.
11. **AGDISTIS**: This approach [61] is a pure entity disambiguation approach (D2KB) based on string similarity measures, an expansion heuristic for labels to cope with co-referencing and the graph-based HITS algorithm. The authors published datasets[25] along with their source code and an API.[26] AGDISTIS can only be used for the D2KB task.
12. **Babelfy**: The core of this approach draws on the use of random walks and a densest subgraph algorithm to tackle the word sense disambiguation and entity linking tasks jointly in a multilingual setting [41] thanks to the BabelNet[27] semantic network [44]. Babelfy has been evaluated using six datasets: three from earlier SemEval tasks [48,43,42], one from a Senseval task [55] and two already used for evaluating AIDA [26,25]. All of them are available online but distributed throughout the Web. Additionally, the authors offer a webservice that is limited to 100 requests per day which are extensible for research purposes [40].[28]
13. **FOX**: FOX [56] is an ensemble-learning based framework for RDF extraction from text based. It makes use of the diversity of NLP algorithms to extract entities with a high precision and a high recall. Moreover, it provides functionality for keyword and relation extraction.
14. **FRED**: In 2015, Gangemi et al. [8] present FRED(*), a novel machine reader based on TagMe. However, FRED extends TagMe with entity typing capabilities.
15. **FREME**: Also in 2015, the EU project FREME publish their e-entity service which is based on Conditional Random Fields for NER while NED is relying on a most-frequent-sense method. That is, candidate entities are chosen based on a sense of commonness within a KB.[29]
16. **`entityclassifier.eu`**: Dojchinovski and Kliegr al. [15] present their approach based on

[21] `http://nerd.eurecom.fr`
[22] `http://s16a.org/kea`
[23] `http://github.com/nopper/wat`
[24] `http://dexter.isti.cnr.it`
[25] `https://github.com/AKSW/n3-collection`
[26] `https://github.com/AKSW/AGDISTIS`
[27] `http://babelnet.org`
[28] `http://babelfy.org`
[29] `https://api-dev.freme-project.eu/doc/api-doc/full.html#/e-Entity`

hypernyms and a Wikipedia-based entity classification system which identifies salient words. The input is transformed to a lower dimensional representation keeping the same quality of output for all sizes of input text.

17. **CETUS**: This approach [53] has been implemented as a baseline for the second task of the OKE challenge 2015. It uses hand crafted patterns to search for entity type information. In a second step, the type hierarchy of the YAGO ontology [33] is used to find a type matching the extracted type information. A second approach called CETUS_FOX uses FOX to retrieve the type of the entity.
18. **xLisa**: Zhang and Rettinger present the x-Lisa annotator [66] which is a three-step pipeline based on cross-lingual Linked Data lexica to harness the multilingual Wikipedia. Based on these lexica, they calculate the mention-candidate-similarity using n-grams. In the third step, x-Lisa constructs an entity-mention graph using the Normalized Google Distance as weights for page rank.
19. **DoSer**: In 2016, Zwickelbauer et al. present DoSer [67], a pure named entity linking approach which is - similiar to AGDISTIS - knowledge-base-agnostic. First, DoSer computes semantic embeddings of entities over one or multiple knowledge bases. Second, given an set of mentions, DoSer calculates possible candidate URIs using existing knowledge base surface forms or additional indexes. Finally, DoSer calculates a personalized page rank using the semantic embeddings of a disambiguation graph constructed from links between possible candidates.
20. **PBOH**: This approach [20] is a pure entity disambiguation approach based on light statistics from the English Wikipedia corpus. The authors developed a probabilistic graphical model using pairwise Markov Random Fields to address the entity linking problem. They show that pairwise co-occurrence statistics of words and entities are enough to obtain comparable or better performance than heavy feature engineered systems. They employ loopy belief propagation to perform inference at test time.
21. **NERFGUN**: The most recent NED system is NERFGUN [22]. This approach reuses ideas from many existing systems for collective named entity linking. First, it uses several indexed based on DBpedia to retrieve a set of candidate entities, such as anchor link texts, `rdfs:labels` and page titles. NERFGUN proposes an undirected probabilistic graphical model based on factor graphs where each factor measures the suitability of the resolution of some mention to a given candidate. The set of features is linearly combined by weights and the inference step is based Markov Chain Monte Carlo models.

Table 3 compares the implemented annotation systems of GERBIL and the BAT-Framework. While AGDISTIS has been in the source code of the BAT-Framework provided by a third-party after publication of Cornolti et al.'s initial work [9] in 2014, GERBIL's community effort led to the implementation of overall 15 new annotators as well as the before mentioned generic NIF-based annotator. The AIDA annotator as well as the "Illinois Wikifier" will not be available in GERBIL since we restrict ourselves to webservices. However, these algorithms can be integrated at any time as soon as their webservices are available.

4.4. Datasets

BAT allowed evaluating the performance of different approaches using the AQUAINT, MSNBC, IITB and the four AIDA/CoNLL datasets (Train A, Train B, Test and Complete). With GERBIL, we include several more datasets that are depicted in Table 4 together with their features. The table shows the huge number of different formats preventing a fast and easy benchmarking of a new annotation system without an intermediate evaluation platform like GERBIL.

GERBIL includes the ACE2004 dataset from Ratinov et al. [49] as well as the dataset of Derczynski [13]. We provide an adapter for the Microposts challenge [51] corpora from 2013 to 2016 each consisting of a test and train dataset as well as an additional third dataset in the years 2015 and 2016. Furthermore, we added the ERD challenge dataset [6] consisting of queries as well as the four GERDAQ datasets [10] for entity recognition and linking in natural language questions. Also, GERBIL includes the Senseval 2 and 3 datasets which took place 2001 and 2004 respectively and activates them for the Entity Recognition task only [16,37]. Moreover, we added the UMBC dataset from 2012 by Finin et al. [19] which was created using crowdsourcing of simple entity annotations over Twitter microposts as well as the WSDM2012/Meij dataset [2] which also describes tweets but these were annotated using only two an-

Table 3

Overview of implemented annotator systems. Brackets indicate the existence of the implementation of the adapter but also the inability to use it in the live system.

		BAT-Framework	GERBIL 1.0.0	GERBIL 1.2.5	Experiment
[39]	Wikipedia Miner	✓	✓	(✓)	A2KB
[49]	Illinois Wikifier	✓	(✓)	✓	A2KB
[36]	Spotlight	✓	✓	✓	OKE Task 1
[26]	AIDA	✓	✓	✓	A2KB
[18]	TagMe 2	✓	✓	✓	A2KB
[51]	NERD-ML		✓	✓	A2KB
[58]	KEA		✓	✓	A2KB
[47]	WAT		✓	✓	A2KB
[7]	Dexter		✓	✓	A2KB
[61]	AGDISTIS	(✓)	✓	✓	D2KB
[41]	Babelfy		✓	✓	A2KB
[56]	FOX			✓	OKE Task 1
[8]	FRED			✓	OKE Task 1
	FREME			✓	OKE Task 1
[15]	`entityclassifier.eu`			✓	A2KB
[53]	CETUS			✓	OKE Task 2
[66]	xLisa			✓	A2KB
[67]	DoSer			✓	D2KB
[20]	PBOH			✓	D2KB
[22]	NERFGUN			✓	D2KB
	NIF-based Annotator		✓	✓	any

notators in 2012. Finally, GERBIL includes the Ritter [50] dataset containing roughly 2.400 tweets annotated with 10 NER classes from Freebase.

Moreover, we capitalize upon the uptake of publicly available, NIF based corpora over the last years [57, 52].[30] To this end, GERBIL implements a Java-based NIF [23] reader and writer module which enables loading arbitrary NIF document collections, as well as the communication to NIF-based webservices. Additionally, we integrated four NIF corpora, i.e., the N^3 RSS-500 and N^3 Reuters-128 dataset,[31] as well as the Spotlight Corpus and the KORE 50 dataset.[32] GERBIL supported the Open Knowledge Extraction Challenge in 2015 and 2016 which led to the integration of the 6 datasets for OKE Task 1 and 6 datasets for OKE Task 2.

The extensibility of the datasets in GERBIL is furthermore ensured by allowing users to upload or use already available NIF datasets from DataHub. However, GERBIL is currently only importing already available datasets. GERBIL will regularly check whether new corpora are available and publish them for benchmarking after a manual quality assurance cycle which ensures their usability for the implemented configuration options. Additionally, users can upload their NIF-corpora directly to GERBIL avoiding their publication in publicly available sources. This option allows for rapid testing of entity annotation systems with closed source or licenced datasets.

GERBIL offers currently 12 state-of-the-art datasets reaching from newswire and twitter to encyclopedic corpora of various amounts of texts and entities. Due to license issues we are only able to provide downloads for 31 of them directly but we provide instructions to obtain the others on our project wiki.[33]

[30] `http://datahub.io/dataset?license_id=cc-by&q=NIF`

[31] `https://github.com/AKSW/n3-collection`

[32] `http://www.yovisto.com/labs/ner-benchmarks/`

[33] The licenses and instructions can be found at `https://github.com/AKSW/gerbil/wiki/Licences-for-datasets`.

Table 4

Datasets, their formats and features. Groups of datasets, e.g., for a single challenge, have been grouped together. A ⋆ indicates various inline or keyfile annotation formats. The experiments follow their definition in Section 4.2

Corpus	Format	Experiment	Topic	|Documents|	Avg. Entity/Doc.	Avg. Words/Doc.
ACE2004	MSNBC	A2KB	news	57	5.37	373.9
AIDA/CoNLL	CoNLL	A2KB	news	1393	25.07	189.7
AQUAINT	⋆	A2KB	news	50	14.54	220.5
Derczynski	TSV	A2KB	tweets	182	1.57	20.8
ERD2014	⋆	A2KB	queries	91	0.65	3.5
GERDAQ	XML	A2KB	queries	992	1.72	3.6
IITB	XML	A2KB	mixed	103	109.22	639.7
KORE 50	NIF/RDF	A2KB	mixed	50	2.88	12.8
Microposts2013	Microposts2013	RT2KB	tweets	4265	1.11	18.8
Microposts2014	Microposts2014	A2KB	tweets	3395	1.50	18.1
Microposts2015	Microposts2015	A2KB	tweets	6025	1.36	16.5
Microposts2016	Microposts2015	A2KB	tweets	9289	1.03	15.7
MSNBC	MSNBC	A2KB	news	20	37.35	543.9
N^3 Reuters-128	NIF/RDF	A2KB	news	128	4.85	123.8
N^3 RSS-500	NIF/RDF	A2KB	RSS-feeds	500	1.00	31.0
OKE 2015 Task 1	NIF/RDF	OKE Task 1	mixed	199	5.11	25.5
OKE 2015 Task 2	NIF/RDF	OKE Task 2	mixed	200	3.06	28.7
OKE 2016 Task 1	NIF/RDF	OKE Task 1	mixed	254	5.52	26.6
OKE 2016 Task 2	NIF/RDF	OKE Task 2	mixed	250	2.83	27.5
Ritter	Ritter	RT2KB	news	2394	0.62	19.4
Senseval 2	XML	ERec	mixed	242	9.86	21.3
Senseval 3	XML	ERec	mixed	352	5.70	14.7
Spotlight Corpus	NIF/RDF	A2KB	news	58	5.69	28.6
UMBC	UMBC	RT2KB	tweets	12973	0.97	17.2
WSDM2012/Meij	TREC	C2KB	tweets	502	1.87	14.4

4.5. Output

GERBIL's main aim is to provide comprehensive, reproducible and publishable experiment results. Hence, GERBIL's experimental output is represented as a table containing the results, as well as embedded JSON-LD[34] RDF data using the RDF DataCube vocabulary [12]. We ensure a detailed description of each component of an experiment as well as machine-readable, interlinkable results following the 5-star Linked Data principles. Moreover, we provide a persistent and time-stamped URL for each experiment that can be used fur publications as it has been done in

RDF DataCube is a vocabulary standard and can be used to represent fine-grained multidimensional, statistical data which is compatible with the Linked SDMX [5] standard. Every GERBIL experiment is modelled as `qb:Dataset` containing the individual runs of the annotators on specific corpora as `qb:Observations`. Each observation features the `qb:Dimensions` experiment type, matching type, annotator, corpus, and time. The evaluation measures offered by GERBIL as well as the error count are expressed as `qb:Measures`. To include further metadata, annotator and corpus dimension properties link *DataID* [3] descriptions of the individual components.

GERBIL uses the recently proposed DataID [3] ontology that combines VoID [1] and DCAT [32] metadata with Prov-O [30] provenance information and ODRL [35] licenses to describe datasets. Besides metadata properties like titles, descriptions and authors, the source files of the open datasets themselves are linked as `dcat:Distributions`, allowing direct access to the evaluation corpora. Furthermore, ODRL license specifications in RDF are linked via `dc:license`, potentially facilitating automatically adjusted processing of licensed data by NLP tools. Li-

[34] `http://www.w3.org/TR/json-ld/`

censes are further specified via `dc:rights`, including citations of the relevant publications.

To describe annotators in a similar fashion, we extended DataID for services. The class `Service`, to be described with the same basic properties as dataset, was introduced. To link an instance of a `Service` to its distribution the `datid:distribution` property was introduced as super property of `dcat:distribution`, i.e., the specific URI the service can be queried at. Furthermore, Services can have a number of `datid:Parameters` and `datid:Configurations`. Datasets can be linked via the `datid:input` or `datid:output` properties.

Offering such detailed and structured experimental results opens new research avenues in terms of tool and dataset diagnostics to increase decision makers' ability to choose the right settings for the right use case. Next to individual configurable experiments, GERBIL offers an overview of recent experiment results belonging to the same experiment and matching type in the form of a table as well as sophisticated visualizations,[35] see Figure 4. This allows for a quick comparison of tools and datasets on recently run experiments without additional computational effort.

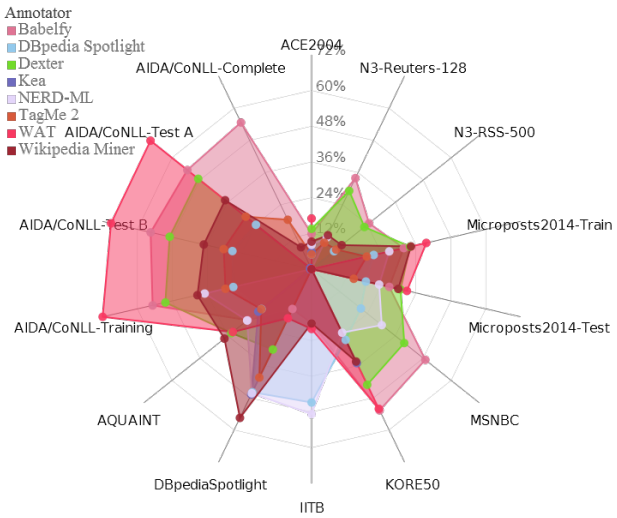

Fig. 4. Example spider diagram of recent A2KB experiments with strong annotation matching derived from our online interface

[35] `http://gerbil.aksw.org/gerbil/overview`

Table 5

Number of tasks executed per annotator. By caching results we did not need to execute 12466 tasks but only 9906. Data taken from 15th February 2015.

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

5. Evaluation

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 20th October 2016, 3.288 experiments were started on the platform containing more than 24.341 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 heavy exploitation of the possibility to test NIF-based webservices, see Table 5. Thus, GERBIL is a powerful evaluation tool for developing new annotators that can be evaluated easily by using the NIF-based interface. Moreover, GERBIL has already been extendedoutside of the core developer team to foster a deeper analysis of annotation systems [64]. Finally, experiment URIs provided by GERBIL are used in over 19 papers using three times the provided stable URIs by the 21st January 2017.

6. Conclusion and Future Work

In this paper, we presented and evaluated GERBIL, a platform for the evaluation of annotation frame-

works. With GERBIL, we aim to push annotation system developers to better quality and wider use of their frameworks. Some of the main contributions of GERBIL include the provision of persistent URLs for reproducibility and archiving. Furthermore, we implemented a generic adaptor for external datasets as well as a generic interface to integrate remote annotator systems. The datasets available for evaluation in the previous benchmarking platforms for annotation was extended by 37 new datasets. Moreover, 15 novel annotators were added to the platform. The presented, web-based frontend allows for several use cases enabling laymen and expert users to perform informed comparisons of semantic annotation tools. The persistent URIs enhance the long term quotation in the field of information extraction. GERBIL is not just a new framework wrapping existing technology. In comparison to earlier frameworks, it extends the state-of-the-art benchmarks by the capability of considering the influence of NIL attributes and the ability of dealing with data sets and annotators that link to different knowledge bases. More information about GERBIL and its source code can be found at the project's website.

In the future, GERBIL will be further enhanced inside the HOBBIT project[36]. GERBIL will be incorporated into a larger benchmarking platform that allows a fair comparison not only of the quality but also of the effectiveness of annotation systems. A first step in that direction is the Open Knowledge Extraction Challenge 2017 which contains tasks that use GERBIL as benchmark measuring the quality of annotation systems executed directly inside the HOBBIT platform to make sure that all participating systems are executed on the same hardware.

Another development is the further support of developers with direct feedback, i.e., showing the annotations that have been marked wrong in the documents. This feature has not been implemented because of licensing issues. However, we think that it would be possible to implement it without license violations for datasets that are publicly available.

Acknowledgments. This work was supported by the German Federal Ministry of Education and Research under the project number 03WKCJ4D and the Eurostars projects DIESEL (E!9367) and QAMEL (E!9725) as well as the European Union's H2020 research and innovation action HOBBIT under the Grant Agreement number 688227.

[36] http://project-hobbit.eu

References

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets with the void vocabulary, 2011. http://www.w3.org/TR/void/.

[2] R. Blanco, G. Ottaviano, and E. Meij. Fast and space-efficient entity linking in queries. In *Proceedings of the eighth ACM international conference on Web search and data mining*, WSDM 2015, 2015.

[3] M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards semantically rich metadata for complex datasets. In *10th International Conference on Semantic Systems 2014*, 2014.

[4] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making sense of microposts (#microposts2014) named entity extraction & linking challenge. In *Proceedings of 4th Workshop on Making Sense of Microposts (#Microposts2014)*, 2014.

[5] S. Capadisli, S. Auer, and A.-C. Ngonga Ngomo. Linked SDMX data. *Semantic Web Journal*, 2013.

[6] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, and K. Wang. ERD 2014: Entity recognition and disambiguation challenge. *SIGIR Forum*, 2014.

[7] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, 2013.

[8] 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. Springer International Publishing, 2015.

[9] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *22nd World Wide Web Conference*, 2013.

[10] M. Cornolti, P. Ferragina, M. Ciaramita, S. Rüd, and H. Schütze. A piggyback system for joint entity mention detection and linking in web queries. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 567–578, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[11] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Conference on Empirical Methods in Natural Language Processing-CoNLL*, 2007.

[12] R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube Vocabulary, 2014. http://www.w3.org/TR/vocab-data-cube/.

[13] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51(2):32 – 49, 2015.

[14] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *LREC*, 2004.

[15] M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-time classification of entities in text with wikipedia. In *Proceedings of the ECMLPKDD'13*, 2013.

[16] P. Edmonds and S. Cotton. Senseval-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating*

Word Sense Disambiguation Systems, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[17] M. V. Erp, G. Rizzo, and R. Troncy. Learning with the web: Spotting named entities on the intersection of NERD and machine learning. In *Proceedings of the Making Sense of Microposts (#MSM2013) Concept Extraction Challenge*, 2013.

[18] P. Ferragina and U. Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE software*, 2012.

[19] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[20] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann. Probabilistic bag-of-hyperlinks model for entity linking. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 927–938, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[21] Y. Gil. Semantic challenges in getting work done, 2014. Invited Talk at ISWC.

[22] S. Hakimov, H. ter Horst, S. Jebbara, M. Hartung, and P. Cimiano. Combining textual and graph-based features for named entity disambiguation using undirected probabilistic graphical models. In *Proceedings of 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2016.

[23] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using Linked Data. In *12th International Semantic Web Conference*, 2013.

[24] 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.

[25] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. KORE: keyphrase overlap relatedness for entity disambiguation. In *Proceedings of CIKM*, 2012.

[26] 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*, 2011.

[27] Y.-m. L. Y.-h. L. W.-h. Z. Hui CHEN, Bao-gang WEI. An easy-to-use evaluation framework for benchmarking entity recognition and disambiguation systems. *Frontiers of Information Technology & Electronic Engineering*, -1(-1), 1998.

[28] P. Jermyn, M. Dixon, and B. J. Read. Preparing clean views of data for data mining. *ERCIM Work. on Database Res*, 1999.

[29] A. Kilgarriff. Senseval: An exercise in evaluating word sense disambiguation programs. *1st LREC*, 1998.

[30] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology, 2013. http://www.w3.org/TR/prov-o/.

[31] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2014.

[32] F. Maali, J. Erickson, and P. Archer. Data Catalog Vocabulary (DCAT), 2014. http://www.w3.org/TR/vocab-dcat/.

[33] F. Mahdisoltani, J. Biega, and F. Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2014.

[34] P. McNamee. Overview of the TAC 2009 knowledge base population track, 2009.

[35] M. McRoberts and V. Rodríguez-Doncel. Open Digital Rights Language (ODRL) Ontology, 2014. http://www.w3.org/ns/odrl/2/.

[36] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[37] R. Mihalcea, T. Chklovski, and A. Kilgarriff. The Senseval-3 English Lexical Sample Task. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004*. Association for Computational Linguistics (ACL), 2004.

[38] P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, editors. *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*. Springer, 2014.

[39] D. Milne and I. H. Witten. Learning to link with wikipedia. In *17th ACM CIKM*, 2008.

[40] A. Moro, F. Cecconi, and R. Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *Proc. of ISWC (P&D)*, 2014.

[41] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *TACL*, 2014.

[42] R. Navigli, D. Jurgens, and D. Vannella. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proceedings of SemEval-2013*, 2013.

[43] R. Navigli, K. C. Litkowski, and O. Hargraves. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. In *Proc. of SemEval-2007*, 2007.

[44] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 2012.

[45] 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.

[46] R. D. Peng. Reproducible research in computational science. *Science (New York, Ny)*, 2011.

[47] F. Piccinno and P. Ferragina. From TagME to WAT: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, 2014.

[48] S. S. Pradhan, E. Loper, D. Dligach, and M. Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proc. of SemEval-2007*, 2007.

[49] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *ACL*, 2011.

[50] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, 2011.

[51] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the extraction and disambiguation of named entities on the semantic web. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.

[52] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. N3 - a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *9th LREC*, 2014.

[53] 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.

[54] M. Rowe, M. Stankovic, and A.-S. Dadzie, editors. *Proceedings, 4th Workshop on Making Sense of Microposts (#Microposts2014): Big things come in small packages, Seoul, Korea, 7th April 2014*, 2014.

[55] B. Snyder and M. Palmer. The English all-words task. In *Proc. of Senseval-3*, pages 41–43, 2004.

[56] R. Speck and A. N. Ngomo. Ensemble learning for named entity recognition. In Mika et al. [38], pages 519–534.

[57] N. Steinmetz, M. Knuth, and H. Sack. Statistical analyses of named entity disambiguation benchmarks. In *1st Workshop on NLP&DBpedia 2013*, 2013.

[58] N. Steinmetz and H. Sack. Semantic multimedia information retrieval based on contextual descriptions. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 382–396. Springer Berlin Heidelberg, 2013.

[59] B. M. Sundheim. Tipster/MUC-5: Information extraction system evaluation. In *Proceedings of the 5th Conference on Message Understanding*, 1993.

[60] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, 2003.

[61] R. Usbeck, A. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. AGDISTIS - graph-based disambiguation of named entities using linked data. In Mika et al. [38], pages 457–471.

[62] R. Usbeck, M. Röder, A. N. 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 annotator benchmarking framework. In A. Gangemi, S. Leonardi, and A. Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1133–1143. ACM, 2015.

[63] 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.

[64] J. Waitelonis, H. Jürges, and H. Sack. Don't compare apples to oranges: Extending GERBIL for a fine grained NEL evaluation. In A. Fensel, A. Zaveri, S. Hellmann, and T. Pellegrini, editors, *Proceedings of the 12th International Conference on Semantic Systems, SEMANTICS 2016, Leipzig, Germany, September 12-15, 2016*, pages 65–72. ACM, 2016.

[65] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[66] L. Zhang and A. Rettinger. X-lisa: Cross-lingual semantic annotation. *PVLDB*, 7(13):1693–1696, 2014.

[67] S. Zwicklbauer, C. Seifert, and M. Granitzer. *DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings*, pages 182–198. Springer International Publishing, Cham, 2016.