GERBIL – General Entity Annotator Benchmarking Framework

Ricardo Usbeck*
Leipzig University, IFI/AKSW
Unister GmbH, Leipzig
usbeck@informatik.uni-leipzig.de

Michael Röder*
Leipzig University, IFI/AKSW
Unister GmbH, Leipzig
roeder@informatik.uni-leipzig.de

Axel-Cyrille Ngonga Ngomo*
Leipzig University, IFI/AKSW
Leipzig (Germany)
ngonga@informatik.uni-leipzig.de

ABSTRACT

The need to bridge between the unstructured data on the Document Web and the structured data on the Web of Data has led to the development of a considerable number of annotation tools. However, these tools are currently still hard to compare since the published evaluation results are calculated on diverse datasets and evaluated based on different measures. We present GERBIL, an evaluation framework for semantic entity annotation. 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. Finally, 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. 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.

1. INTRODUCTION

The implementation of the original vision behind the Semantic Web demands the development of approaches and frameworks for the seamless extraction of structured data from text. 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 [13, 17, 24, 25, 27, 32, 35, 41, 44], 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 preparing data for experiments [14, 18, 31]. 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 [7, 39]. 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.

The insights above have led to a movement towards the creation of frameworks to ease the evaluation of solutions that address the same annotation problem [5, 7]. In this paper, we present GERBIL – a general entity annotator benchmark –, a community-driven effort to enable the continuous evaluation of annotation tools. GERBIL is an open-source and extensible framework that allows evaluating tools against (currently) 9 different annotators on 11 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. While the evaluation core of GERBIL is based on the BAT-framework

*Further co-authors can be found at the end of the article.

[7], our approach goes beyond the state of the art in several respects:

- 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.
- 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.
- 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.
- The results of GERBIL are published in a *machine-readable format.* In particular, our use of DataID [2] and DataCube [9] 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 reconstructable. By these means, we also tackle the problem of *reproducibility.*
- 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.
- 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 principles 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 6 additional annotators not evaluated against each other in previous works (e.g., [7]).
- **Easy integration of datasets**: We also provide means to gather datasets for evaluation directly from data services such as DataHub.[1] In particular, we added 6 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[2] 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.

In the rest of this paper, we present and evaluate GERBIL. We begin by giving an overview of related work. Thereafter, we present the GERBIL framework. We focus in particular on how annotators and datasets can be added to GERBIL and give a short overview of the annotators and tools that are currently included in the framework. We then present an evaluation of the framework that aims to quantify the effort necessary to include novel annotators and datasets to the framework. 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`.

2. 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 hand full of frameworks.

The BAT-framework [7] is designed to facilitate the benchmarking of named entity recognition (NER), named entity disambiguation (NED) – also known as linking (NEL) – 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. [35] 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 [3] for NED. The authors propose a common schema, named the NERD ontology[3], 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.

Over the course of the last 25 years several challenges, workshops and conference dedicated themselves to the compara-

[1] `http://datahub.io`

[2] Available at `http://gerbil.aksw.org`.

[3] `http://nerd.eurecom.fr/ontology`

ble evaluation of information extraction (IE) systems. Starting in 1993, the Message Understanding Conference (MUC) introduced a first systematic comparison of information extraction approaches [42]. 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 [43]. In addition, the Automatic Content Extraction (ACE) challenge [10], 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) [22] where mainly linguistic-based approaches are published. The Senseval challenge, originally concerned with classical NLP disciplines, has wided it focus in 2007 and changed its name to SemEval to account for the recently recognized impact of semantic technologies [19]. The Making Sense of Microposts workshop series (MSM) established in 2013 an entity recognition and in 2014 an entity linking challenge thereby focusing on tweets and microposts [37]. In 2014, Carmel et al. [5] 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 [35] in several dimensions to enhance reproducibility, diagnostics and publishability of entity annotation systems. In particular, we provide 6 additional datasets and 6 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 [15] interfaces, which are designed to simplify the exchange of data and binding of services.

3. THE GERBIL FRAMEWORK

3.1 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 metric 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.

3.2 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.

3.2.1 Experiment types

An experiment type defines the way used to solve a certain problem when extracting information. Cornolti et al.'s [7] BAT-framework offers six different experiment types, namely (scored) annotation (S/A2KB), disambiguation (D2KB) – also known as linking –, (scored respectively ranked) concept annotation (S/R/C2KB) of texts. In [35], 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 reuses the six experiments provided by the BAT-framework and extends them 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 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 Ricardo does not have a URI in these reference datasets. Our framework extends the experiments types of [7] 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 six types of experiments:

1. **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.

2. **A2KB**: This task is the classical NER/D task, thus an extension of the D2KB task. Here two functions are to be found. First, 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. The aim of the second step is then to match the results of ex to entities from $K \cup \{NIL\}$ by devising a function a as in the D2KB task.

3. **Sa2KB**: Sa2KB is an extension of A2KB where the scores $c_i \in [0, 1]$ of the mentions detected by the approach are taken into consideration. These scores are then used during the evaluation.

4. **C2KB**: The concept tagging task C2KB aims to detect entities when given a document. Formally, the tagging

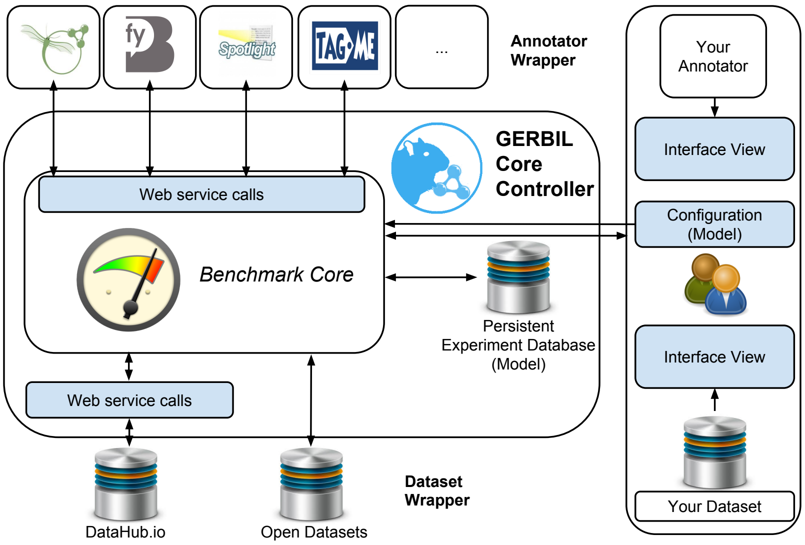

Figure 1: Overview of GERBIL's abstract architecture. Interfaces to users and providers of datasets and annotators are marked in blue.

function *tag* simply returns a subset of K for each input document d.

5. **Sc2KB**: This task is an extension of C2KB where the tagging function returns a subset of $K \times [0,1]$ for each input document d.

6. **Rc2KB**: In this particular extension of C2KB, the tagging function returns a sorted list of resources from K, i.e., an element of K^*, where $K^* = \bigcup_{i=0}^{\infty} K^i$.

With this extension, our framework can now deal with gold standard datasets and annotators that link to any knowledge base, e.g., DBpedia, BabelNet [30] etc., as long as the necessary identifiers are URIs. We were thus able to implement 6 new gold standard datasets usable for each experiment type, cf. Section 3.3, and 6 new annotators linking entities to any knowledge base instead of solely Wikipedia like in previous works, cf. Section 3.2.4. With this extensible interface, GERBIL can be extended to deal with supplementary experiment types, e.g., entity salience [7], entity detection [39], typing [35], word sense disambiguation (WSD) [27] and relation extraction [39]. These categories of experiment types will be added to GERBIL in next versions.

3.2.2 Matching

A matching defines which conditions the result of an annotator has to fulfill to be a correct result. In case of existing redirections, we assume an implicit matching function to account for the many-to-one relation [7]. The first matching type M used for the C2KB, Rc2KB and Sc2KB experiments is the *strong entity matching*. Here, each mention is mapped to an entity of the knowledge base K via a matching function f with $f(m) \in K \cup \{NIL\}$. Following this matching, a single entity mention $m = (s,l,d,c)$ returned by the annotator is correct iff it matches exactly with one of the entity mentions $m' = (s',l',d,c')$ in the gold standard $G(d)$ of d [7]. Formally,

$$M(m,G) = \begin{cases} 1 & \text{iff } \exists m' \in G, f(m) = f(m'), \\ 0 & \text{else.} \end{cases} \tag{1}$$

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

$$M_e(m,G) = \begin{cases} 1 & \text{iff } \exists m' \in G : f(m) = f(m') \wedge s = s' \wedge \\ & l = l', \\ 0 & \text{else.} \end{cases} \tag{2}$$

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* 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.

$$M_w(m,G) = \begin{cases} 1 & \text{iff } \exists m' \in G, f(m) = f(m') \wedge (\\ & \quad (s \le s' \wedge (s+l) \le (s'+l')) \\ & \quad \vee (s \ge s' \wedge (s+l) \ge (s'+l')) \\ & \quad \vee (s \le s' \wedge (s+l) \ge (s'+l')) \\ & \quad \vee (s \ge s' \wedge (s+l) \le (s'+l'))) \\ 0 & \text{else.} \end{cases} \tag{3}$$

3.2.3 Metrics

Currently, GERBIL offers 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. At the moment, those measures ignore NIL annotations, i.e., if a gold standard dataset contains entities that are not contained in the target knowledge base K and an annotator detects the entity and links it to any URI, emerging novel URI or NIL, this will always result in a false-positive evaluation. To alleviate this problem, GERBIL allows adding additional measures to evaluate the results of annotators regarding the heterogeneous landscape of gold standard datasets.

3.2.4 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. Since GERBIL is based on the BAT-framework, annotators of this framework can be added to GERBIL easily. Due to the community effort behind GERBIL, we could raise the number of published annotators from 5 to 9. We investigated the effort to implement a BAT-framework adapter in contrast to evaluation efforts done without a structured evaluation framework in Section 4.

2. **NIF-based Services**: GERBIL implements means to understand NIF-based [15] communication over webservice 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.[4] 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[5]. GERBIL does not store these connections in terms of API keys or URLs but still offers the opportunity of persistent experiment results.

Currently, GERBIL offers 9 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 [8]. 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[6] 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 [25] 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[7]. They provide the source code for their approach as well as a webservice[8] which is available in GERBIL.

3. **Illinois Wikifier**: In 2011, [34] 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[9] as well as their software ”Illionois Wikifier”[10] online. Since ”Illionois 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 [24] was published in 2011, this framework combines NER and NED approach based upon DBpedia[11]. Based on a vector-space representation of entities and using the cosine similarity, this approach has a public (NIF-based) webservice[12] as well as its online available evaluation dataset[13].

5. **TagMe 2**: TagMe 2 [13] was publised 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 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[14] as well as their datasets[15] online. The source

[4] 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.

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

[6] `http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/`

[7] `http://www.nist.gov/tac/data/data_desc.html#AQUAINT`

[8] `http://wikipedia-miner.cms.waikato.ac.nz/`

[9] `http://cogcomp.cs.illinois.edu/page/resource_view/4`

[10] `http://cogcomp.cs.illinois.edu/page/software_view/33`

[11] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Known-uses`

[12] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service`

[13] `http://wiki.dbpedia.org/spotlight/isemantics2011/evaluation`

[14] `http://tagme.di.unipi.it/`

[15] `http://acube.di.unipi.it/tagme-dataset/`

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.

6. **AIDA**: The AIDA approach [17] relies on coherence graph building and dense subgraph algorithms and is based on the YAGO2[16] 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 [43], GERBIL will not use the webservice since it is not stable enough for regular replication purposes.[17]

7. **NERD-ML**: In 2013, [11] 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 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 [35], improved the classification task by redesigning the selection of the features, and they proposed experiments on both microposts and newswire domains. NERD-ML has a public webservice which is part of GERBIL[18].

8. **KEA NER/NED**: This approach is the successor of the approach introduced in [41] 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[19].

9. **WAT**: WAT is the successor of TagME [13].[20] 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. [13]). 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. **AGDISTIS**: This approach [44] 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[21] along with their source code and an API[22]. AGDISTIS can only be used for the D2KB task.

11. **Babelfy**: The core of this approach lies in the use of random walks and a densest subgraph algorithm to tackle the word sense disambiguation and entity linking tasks in a multilingual setting [27] thanks to the BabelNet semantic network [30]. Babelfy has been evaluated using six datasets: three from earlier SemEval tasks [33, 29, 28], one from a Senseval task [38] and two already used for evaluating AIDA [17, 16]. All of them are available online but distributed throughout the web. Additionally, the authors offer a webservice limited to 100 requests per day which are extensible for research purposes[23] [26].

12. **Dexter**: This approach [6] 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.

Table 1 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 [7] in 2014, GERBIL's community effort led to the implementation of overall 6 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.

3.3 Datasets

Table 2 shows the heterogeneity of datasets used for prior evaluations while Table 3 presents an overview of the datasets that were used to evaluate some well-known entity annotators in previous works. These tables make clear that the numbers and types of used datasets varies a lot, thus preventing a fast comparison of annotation systems.

BAT allows evaluating the performance of different approaches using five datasets, namely AQUAINT, MSNBC, IITB, Meij and AIDA/CoNLL. With GERBIL, we activate one more dataset already implemented by the authors, namely ACE2004 from Ratinov et al. [34]. Furthermore, we implemented a dataset wrapper for the Microposts2014 corpus which has been used to evaluate NERD-ML [35]. The

[16] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

[17] `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/`

[18] `http://nerd.eurecom.fr/`

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

[20] `http://github.com/nopper/wat`

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

[22] `https://github.com/AKSW/AGDISTIS`

[23] `http://babelfy.org`

[24] `http://dexter.isti.cnr.it`

	Year	ACE	Wiki	Aquaint	MSNBC	IITB	Meij	AIDA/CoNLL	N^3 collection	KORE 50	Wiki-Disamb30	Wiki-Annot30	Spotlight Corpus	SemEval-2013 task 12	SemEval-2007 task 7	SemEval-2007 task 17	Senseval-3	NIF-based corpus	Microposts2014	Software available?	Webservice available?
Cucerzan	2007				✓																
Wikipedia Miner	2008			✓*																	✓
Illionois Wikifier	2011	✓	✓	✓*	✓															✓	
Spotlight	2011												✓							✓	✓
AIDA	2011							✓												✓	✓**
TagMe 2	2012										✓	✓								✓	✓
Dexter	2013																			✓	✓
KEA	2013																				✓
WAT	2013																			✓	✓
AGDISTIS	2014			✓	✓	✓	✓	✓	✓	✓			✓							✓	✓
Babelfy	2014							✓		✓				✓	✓	✓	✓				✓
NERD-ML	2014							✓											✓	✓	✓
BAT-Framework	2013	✓	✓	✓	✓	✓	✓	✓*												✓	
NERD Framework	2014						✓	✓											✓	✓	✓
GERBIL	2014	✓	✓	✓	✓	✓	✓	✓*	✓	✓			✓					✓	✓	✓	✓

Table 3: Comparison of annotators and datasets with indication whether software or datasets respectively webservices are available for reproduction. $*$ indicates that only a subset has been used to evaluate this annotator. $**$ indicate that the webservice is not meant to be used within scientific evaluations due to unstable backends.

dataset itself was introduced in 2014 [3] and consists of 3500 tweets especially related to event data. Moreover, we capitalize upon the uptake of publicly available, NIF based corpora over the last years [40, 36][25]. To this end, GERBIL implements a Java-based NIF [15] 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 RSS-500 and reuters-128 dataset[26], as well as the Spotlight Corpus and the KORE 50 dataset[27].

The extensibility of the datasets in GERBIL is furthermore ensured by allowing users to upload or use already available NIF datasets from DataHub. 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.

Some of the datasets shown in Table 3 are either not yet implemented due to size and server load limitations, i.e., Wiki-Disamb30 and Wiki-Annot30, or due their original experiment type. In particular, the Senseval-3 as well as the different SemEval datasets demand as experiment type word sense disambiguation and thereby linking to BabelNet or Wordnet [12], which is not yet covered in GERBIL. Still, GERBIL offers currently 11 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 9 of them directly but we provide instructions to obtain the others on our project wiki.

Table 4 depicts the features of the current datasets available in GERBIL. These provide a broad evaluation ground leveraging the possibility for sophisticated tool diagnostics.

3.4 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[28] RDF data using the RDF DataCube vocabulary [9]. 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

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

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

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

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

Corpus	Topic	|Documents|	Avg. Entity/Doc.
ACE2004	news	57	4.44
AIDA/CoNLL	news	1393	19.97
Aquaint	news	50	14.54
IITB	mixed	103	109.22
KORE 50	mixed	50	2.86
Meij	tweets	502	1.62
Microposts2014	tweets	3505	0.65
MSNBC	news	20	32.50
N^3 Reuters-128	news	128	4.85
N^3 RSS-500	RSS-feeds	500	0.99
Spotlight Corpus	news	58	5.69

Table 4: Features of the datasets and their documents.

		BAT-Framework	GERBIL	Experiment
[25]	Wikipedia Miner	✓	✓	SA2KB
[34]	Illionois Wikifier	✓	(✓)	SA2KB
[24]	Spotlight	✓	✓	SA2KB
[13]	TagMe 2	✓	✓	SA2KB
[17]	AIDA	✓	(✓)	SA2KB
[41]	KEA		✓	SA2KB
[32]	WAT		✓	SA2KB
[44]	AGDISTIS	(✓)	✓	D2KB
[27]	Babelfy		✓	SA2KB
[35]	NERD-ML		✓	SA2KB
[6]	Dexter		✓	SA2KB
	NIF-based Annotator		✓	any

Table 1: 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.

Dataset	Format	Experiment
ACE2004	MSNBC	Sa2KB
Wiki	⋆	Sa2W
Aquaint	⋆	Sa2KB
MSNBC	MSNBC	Sa2KB
IITB	XML	Sa2KB
Meij	TREC	Rc2W
AIDA/CoNLL	CoNLL	Sa2KB
N^3 collection	NIF/RDF	Sa2KB
KORE 50	NIF/RDF	Sa2KB
Wiki-Disamb30	tab-separated	Sa2KB
Wiki-Annot30	tab-separated	Sa2KB
Spotlight Corpus	NIF/RDF	Sa2KB
SemEval-2013 task 12	XML/⋆	WSD/Sa2KB
SemEval-2007 task 7	XML/⋆	WSD
SemEval-2007 task 17	XML/⋆	WSD
Senseval-3	XML/⋆	WSD
Microposts2014	Microposts2014	Sa2KB

Table 2: Datasets and their formats. A ⋆ indicates various inline or keyfile annotation formats. The experiments follow their definition in Section 3.2

persistent and time-stamped URL for each experiment, see Table 5.

RDF DataCube is a vocabulary standard and can be used to represent fine-grained multidimensional, statistical data which is compatible with the Linked SDMX [4] 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 six 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* [2] descriptions of the individual components.

GERBIL uses the recently proposed DataID [2] ontology that combines VoID [1] and DCAT [21] metadata with Prov-O [20] provenance information and ODRL [23] 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. Licenses 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 `datid:input` or `datid:output`.

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

Annotator	Dataset	F1-micro
DBpedia Spotlight	IITB	0.444
Babelfy	IITB	0.377
NERD-ML	IITB	0.488
WAT	IITB	0.202
DBpedia Spotlight	KORE50	0.265
Babelfy	KORE50	0.476
NERD-ML	KORE50	0.238
WAT	KORE50	0.523

Table 5: Results of an example experiment. It is accessible at `http://gerbil.aksw.org/gerbil/experiment?id=201411100001`

matching type in the form of a table as well as sophisticated visualizations[29], see Figure 2. This allows for a quick comparison of tools and datasets on recently run experiments without additional computational effort.

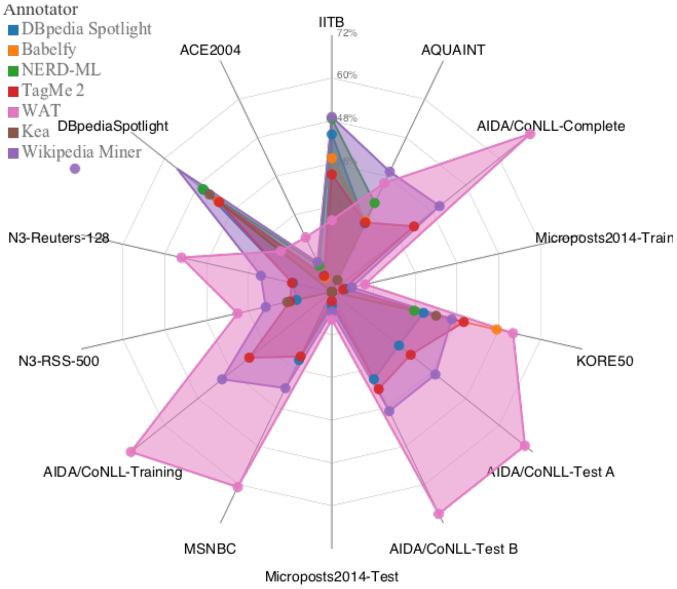

Figure 2: Example spider diagram of recent A2KB experiments with strong annotation matching derived from our online interface

4. EVALUATION

To ensure the practicability and convenience of the GERBIL framework, we investigated the effort needed to use GERBIL for the evaluation of novel annotators. To achieve this goal, we surveyed the workload necessary to implement a novel annotator into GERBIL compared to the implementation into previous diverse frameworks.

Our survey comprised five developers with expert-level programming skills in Java. Each developer was asked to evaluate how much time he/she needed to write the code necessary to evaluate his/her framework on a new dataset.

Overall, the developers reported that they needed between 1 and 4 hours to achieve this goal (4x 1-2h, 1x 3-4h), see Figure 3. Importantly, all developers reported that they needed either the same or even less time to integrate their annotator into GERBIL. This result in itself is of high practical significance as it means that by using GERBIL, developers can evaluate on (currently) 11 datasets using the same effort they needed for 1, which is a gain of more than 1100%. Moreover, all developers reported they felt comfortable—4 points on average on a 5-point Likert scale between very uncomfortable (1) and very comfortable (5)—implementing the annotator in GERBIL. Further developers were invited to complete the survey, which is available at our project website. Even though small, this evaluation suggests that implementing against GERBIL does not lead to any overhead. On the contrary, GERBIL significantly improves the time-to-evaluation by offering means to benchmark and compare against other annotators respectively datasets within the same effort frame previously required to evaluate on a single dataset.

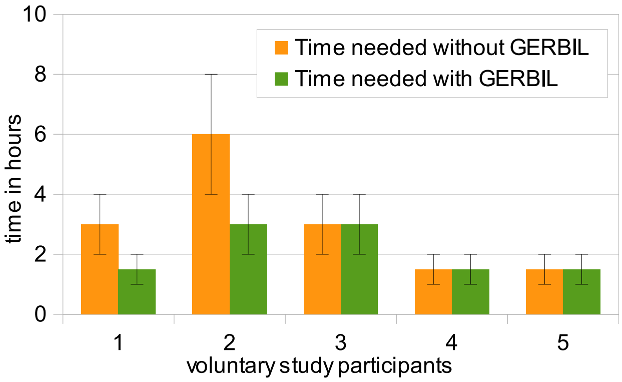

Figure 3: Comparison of effort needed to implement an adapter for an annotation system with and without GERBIL.

An interesting side-effect of having all these frameworks and datasets in a central framework is that we can now benchmark the different frameworks with respect to their runtimes within exactly the same experimental settings. These results are of practical concern for end users of annotation frameworks as they are most commonly interested in both the runtime and the quality of solutions. For example, we evaluated the runtimes of the different approaches in GERBIL for the A2KB experiment type on the MSNBC dataset. The results of this experiment are shown in Figure 4.

5. CONCLUSION AND FUTURE WORK

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

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

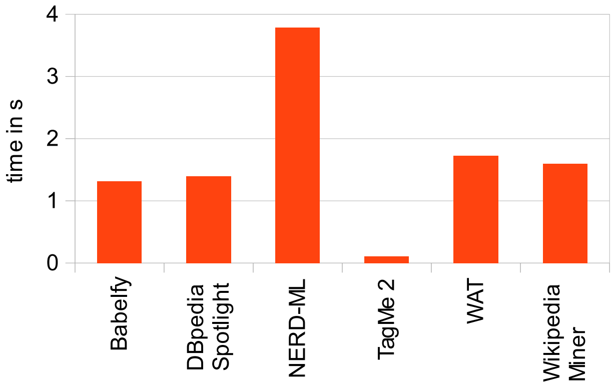

Figure 4: Runtime per document of different approaches in GERBIL for the A2KB experiment type on the MSNBC dataset.

extended by 6 new datasets. Moreover, 6 novel annotators were added to the platform. The evaluation of our framework by contributors suggests that adding an annotator to GERBIL demands 1 to 2 hours of work. Hence, while keeping the implementation effort previously required to evaluate on a single dataset, we allow developers to evaluate on (currently) 11 times more datasets. 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 enhances 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.

While developing GERBIL, we spotted several flaws in the formal model underlying previous benchmarking frameworks which we aim to tackle in the future. For example, the formal specification underlying current benchmarking frameworks for annotation does not allow using the scores assigned by the annotators for their results. To address this problem, we aim to develop/implement novel measures into GERBIL that make use of scores (e.g., Mean Reciprocal Rank). Moreover, partial results are not considered within the evaluation. For example, during the disambiguation task, named entities without Wikipedia URIs are not considered. This has a significant impact of the number of true and false positives and thus on the performance of some tools. Furthermore, certain tasks seem to be too coarse. For example, we will consider splitting the Sa2KB and the A2KB tasks into two subtasks: The first subtask would measure how well tools perform at finding named entities inside the text (NER task) while the second would evaluate how well tools disambiguate those named entities which have been found correctly (similar to the D2KB task). In the future, we also plan to provide information about the point in time since when an annotator is stable, i.e., the algorithm underlying the webservice has not changed.

Acknowledgments. Parts of this work were supported by the ESF and the Free State of Saxony, the FP7 projects GeoKnow (GA No. 318159), LIDER (GA No. 610782), and LinkedTV (GA No. 287911).

6. ADDITIONAL AUTHORS

Ciro Baron (Leipzig University, Germany)
Andreas Both (R&D, Unister GmbH, Germany)
Martin Brümmer (Leipzig University, Germany)
Diego Ceccarelli (Unversity of Pisa, Italy)
Marco Cornolti (University of Pisa, Italy)
Didier Cherix (R&D, Unister GmbH, Germany)
Bernd Eickmann (R&D, Unister GmbH, Germany)
Paolo Ferragina (University of Pisa, Italy)
Christiane Lemke (R&D, Unister GmbH, Germany)
Andrea Moro (Sapienza University of Rome, Italy)
Roberto Navigli (Sapienza University of Rome, Italy)
Francesco Piccinno (University of Pisa, Italy)
Giuseppe Rizzo (EURECOM, France)
Harald Sack (HPI Potsdam, Germany)
René Speck (Institute for Applied Informatics, Germany)
Raphaël Troncy (EURECOM, France)
Jörg Waitelonis (HPI Potsdam, Germany)
Lars Wesemann (R&D, Unister GmbH, Germany)

7. 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] 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.

[3] 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.

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

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

[6] 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.

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

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

[9] R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube Vocabulary, 2014.

http://www.w3.org/TR/vocab-data-cube/.

[10] 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.

[11] 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.

[12] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

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

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

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

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

[17] 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.

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

[19] A. Kilgarri. Senseval: An exercise in evaluating word sense disambiguation programs. *Proc. of the first international conference on language resources and evaluation*, 1998.

[20] 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/.

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

[22] P. McNamee. Overview of the tac 2009 knowledge base population track. 2009.

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

[24] 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.

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

[26] A. Moro, F. Cecconi, and R. Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 13th Internation Conference on Semantic Web (P&D)*, 2014.

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

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

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

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

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

[32] 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.

[33] 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*, pages 87–92. Association for Computational Linguistics, 2007.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] R. Speck and A.-C. N. Ngomo. Ensemble learning for named entity recognition. In *The Semantic Web – ISWC 2014*. 2014.

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

[41] 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*. 2013.

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

[43] 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*.

[44] R. Usbeck, A.-C. Ngonga Ngomo, M. Röder, D. Gerber, S. Coelho, S. Auer, and A. Both. Agdistis - graph-based disambiguation of named entities using linked data. In *The Semantic Web – ISWC 2014*. 2014.