Evaluating Entity Annotators Using GERBIL

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

University of Leipzig, Germany
{usbeck,roeder,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 hard to compare due to the diversity of data sets and measures used for evaluation. We will demonstrate GERBIL, an evaluation framework for semantic entity annotation that provides developers, end users and researchers with easy-to-use interfaces for the agile, fine-grained and uniform evaluation of annotation tools on 11 different data sets within 6 different experimental settings on 6 different measures.

1 Introduction

The need for extracting structured data from text has led to the development of a large number of tools dedicated to the extraction of structured data from unstructured data (see [4] for an overview). In this demo, we present GERBIL, a framework for the evaluation of entity annotation frameworks. GERBIL provides a GUI that allows (1) configuring and running experiments, (2) assigning persistent URLs to experiments (better reproducibility and archiving), (3) exporting the results of the experiments in human- and machine-readable formats as well as (4) displaying the results w.r.t. the data sets and the features of the data sets on which the experiments were performed.

GERBIL is an open-source and extensible framework that allows evaluating tools against (currently) 9 different annotators on 11 sets different data sets within 6 different experiment types. To ensure that our framework is useful to both end users and tool developers, its architecture and interface were designed to allow (1) the easy integration of annotators through REST services, (2) the easy integration of data sets via DataHub[1], file uploads or direct source code integration, (3) the addition of new performance measures, (4) the provision of diagnostics for tool developers and (5) the portability of results. More information on GERBIL as well as a link to the online demo can be found at the project webpage at `http://gerbil.aksw.org`.

2 GERBIL in a nutshell

An overview of GERBIL's architecture is given in Figure 1. Based on this architecture, we will explain the features that we will present in the demonstration of the GERBIL framework.

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

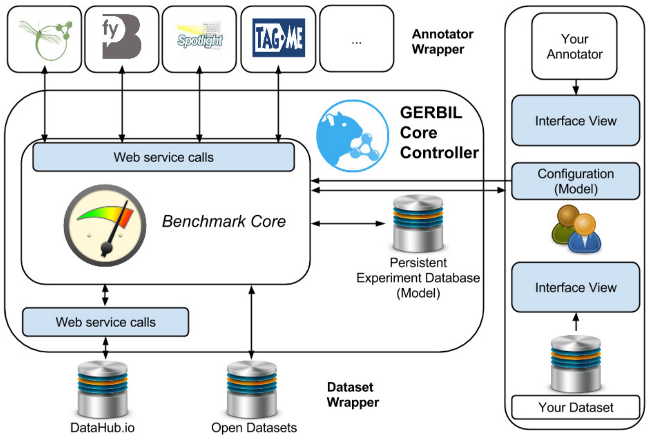
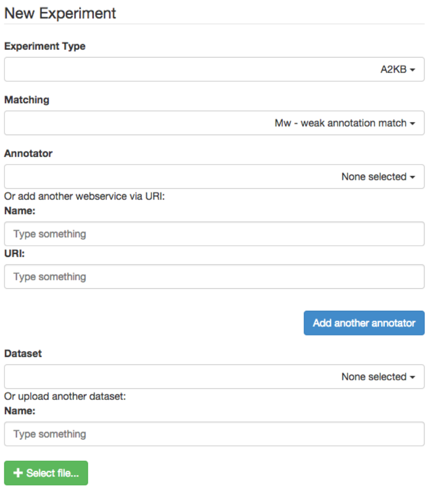


Fig. 1: Overview of GERBIL's abstract architecture. Interfaces to users and providers of data sets and annotators are marked in blue.



New Experiment

Experiment Type

A2KB ▾

Matching

Mw - weak annotation match ▾

Annotator

None selected ▾

Or add another webservice via URI:

Name:

Type something

URI:

Type something

Add another annotator

Dataset

None selected ▾

Or upload another dataset:

Name:

Type something

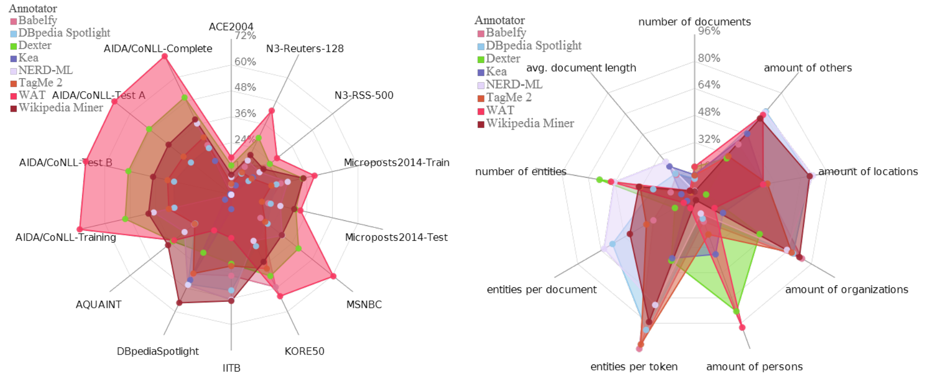
+ Select file...

Fig. 3: Experiment configuration screen.

Feature 1: Experiment types. An experiment type defines the way used to solve a certain problem when extracting information. GERBIL extends the six experiments types provided by the BAT framework [1] (including entity recognition and disambiguation). With this extension, our framework can deal with gold standard data sets and annotators that link to any knowledge base, e.g., DBpedia, BabelNet [3] etc., as long as the necessary identifiers are URIs. During the demo, we will show how users can select the type of experiments in the interface (see Figure 3) and explain the different types of experiments.

Feature 2: Matchings. GERBIL offers three types of matching between a gold standard and the results of annotation systems: a *strong entity matching* for URLs, as well as a *strong* and a *weak annotation matching* for entities. The selection and an explanation of the types of matching for given experiments will be part of the demo (see Figure 3).

Feature 3: Metrics. Currently, GERBIL offers six measures subdivided into two groups: the micro- and the macro-group of precision, recall and f-measure. As shown in Figure 2(a), these results are displayed using interactive spider diagrams that allow the user to easily (1) get an overview of the performance of single tools, (2) compare tools



(a) Example spider diagram of recent A2KB experiments with weak annotation matching.

(b) Spider diagram of correlations between annotation results and data set features.

Fig. 2: Spider diagrams generated by the GERBIL interface.

with each other and (3) gather information on the performance on tools on particular data sets. We will show how to interact with our spider diagrams during the demo.

Feature 4: Diagnostics. An important novel feature of our interface is that it displays the correlation between the features of data sets and the performance of tools (see Figure 2(b)). By these means, we ensure that developers can easily gain an overview of the performance of tools w.r.t. a set of features and thus detect possible areas of improvement for future work.

Feature 5: Annotators. The main goal of GERBIL is to simplify the comparison of novel and existing entity annotation systems in a comprehensive and reproducible way. Therefore, GERBIL offers several ways to implement novel entity annotation frameworks. We will show how to integrate annotators into GERBIL by using a Java adapter as well as a *NIF-based Service* [2].Currently, GERBIL offers 9 entity annotation systems with a variety of features, capabilities and experiments out-of-the-box, including Illinois Wikifier, DBpedia Spotlight, TagMe, AIDA, KEA, WAT, AGDISTIS, Babelfy, NERD-ML and Dexter [4].

Feature 6: Data sets. Table 1 shows the 11 sets data sets available via GERBIL. Thank to the large number of formats, topics and features of the datasets, GERBIL allows carrying out diverse experiments. During the demo, we will show how to add more data sets to GERBIL.

Feature 7: 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[2] RDF data. During the demo, we will show the output generated by GERBIL for the different experiments implemented and show how the RDF results can be used for the sake of archiving results. Moreover, we will show how to retrieve experimental results using the permanent URI generated by GERBIL.

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

Table 1: Features of the data sets and their documents.

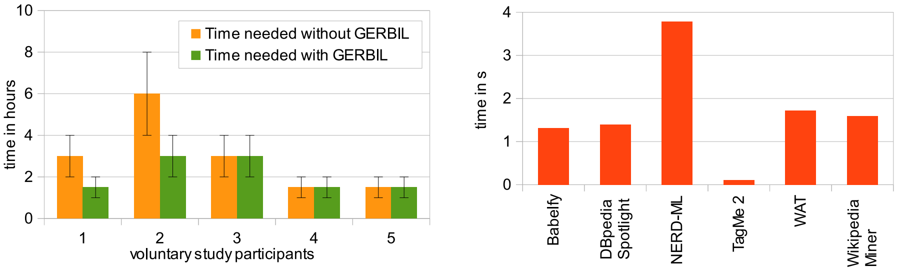
Corpus	Topic	Format	Experiment	Size	Avg. Entity/Doc.
ACE2004	news	MSNBC	Sa2KB	57	4.44
AIDA/CoNLL	news	CoNLL	Sa2KB	1393	19.97
Aquaint	news	-	Sa2KB	50	14.54
IITB	mixed	XML	Sa2KB	103	109.22
KORE 50	mixed	NIF/RDF	Sa2KB	50	2.86
Meij	tweets	TREC	Rc2KB	502	1.62
Microposts2014	tweets	-	Sa2KB	3505	0.65
MSNBC	news	MSNBC	Sa2KB	20	32.50
N^3 Reuters-128	news	NIF/RDF	Sa2KB	128	4.85
N^3 RSS-500	RSS-feeds	NIF/RDF	Sa2KB	500	0.99
Spotlight Corpus	news	NIF/RDF	Sa2KB	58	5.69

3 Evaluation

To ensure that GERBIL can be used in practical settings, 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 data set. Further details pertaining to this evaluation are reported in the research paper to this demo [4].

Overall, the developers reported that they needed between 1 and 4 hours to achieve this goal (4x 1-2h, 1x 3-4h), see Figure 4(a). 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 sets data sets 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. Even though small, this evaluation suggests that implementing against GERBIL does not lead to any overhead. Furthermore, the interviewed developers represent a majority of the active research and development community in the are of entity annotation systems.

An interesting side-effect of having all these frameworks and data sets in a central framework is that we can now benchmark the different frameworks with respect to their runtimes within exactly the same experimental settings. For example, we evaluated the runtimes of the different approaches in GERBIL for the A2KB experiment type on the MSNBC data set, see Figure 4(b).



(a) Comparison of effort needed to implement an adapter for an annotation system.

(b) Runtime per document for the A2KB experiment type on the MSNBC data set.

Fig. 4: Overview of GERBIL evaluation results.

4 Conclusion and Future Work

In this paper, we presented a demo for GERBIL, a platform for the evaluation of annotation frameworks. We presented the different features that make the GERBIL interface easy to use and informative both for end users and developers. With GERBIL, we aim to push annotation system developers to better quality and wider use of their frameworks as well as include the provision of persistent URLs for reproducibility and archiving. GERBIL 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. In future work, we aim to provide a new theory for evaluating annotation systems and display this information in the GERBIL interface.

Acknowledgments. Parts of this work were supported by the FP7 project GeoKnow (GA No. 318159) and the BMWi project SAKE (GA No. 01MD15006E).

References

1. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *22nd World Wide Web Conference*, 2013.
2. S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using Linked Data. In *12th International Semantic Web Conference*, 2013.
3. R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 2012.
4. R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*, 2015.