Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance

Fabian Odoni[a], Philipp Kuntschik[a], Adrian M.P. Braşoveanu[a,b], Albert Weichselbraun[a]

[a]*Swiss Institute for Information Science, University of Applied Sciences Chur, Pulvermühlestrasse 57, 7000 Chur*
[b]*MODUL Technology GmbH, Am Kahlenberg 1, 1190 Vienna, Austria*

Abstract

Rigorous evaluations and analyses of evaluation results are key towards improving Named Entity Linking systems. Nevertheless, most current evaluation tools are focused on benchmarking and comparative evaluations. Therefore, they only provide aggregated statistics such as precision, recall and F1-measure to assess system performance and no means for conducting detailed analyses up to the level of individual annotations.

This paper addresses the need for transparent benchmarking and fine-grained error analysis by introducing Orbis, an extensible framework that supports drill-down analysis, multiple annotation tasks and resource versioning. Orbis complements approaches like those deployed through the GERBIL and TAC KBP tools and helps developers to better understand and address shortcomings in their Named Entity Linking tools.

We present three uses cases in order to demonstrate the usefulness of Orbis for both research and production systems: (i) improving Named Entity Linking tools; (ii) detecting gold standard errors; and (iii) performing Named Entity Linking evaluations with multiple versions of the included resources.

Keywords: Named Entity Linking; Evaluation; Drill-Down Analysis; Resource Versioning; Linked Data Quality;

1. Introduction: The Principles of Transparent Benchmarking

Gold standards, Named Entity Linking (NEL) challenges and rigorous evaluations are key components towards improving Named Entity Linking performance. However, during the years we have noticed that NEL development can be inefficient if evaluations are not followed by drill-down analysis that break errors into classes and investigate

* Corresponding author. Tel.: +41 81 286 38 25.
E-mail address: adrian.brasoveanu@htwchur.ch

means for addressing them. Such analyses often also reveal flaws in gold standard datasets which might have been caused by annotation errors, Knowledge Base (KB) errors, and use of different annotation guidelines. While powerful tools such as GERBIL and TAC KBP support researchers in determining aggregated performance metrics, performing comprehensive drill-down analysis or detailed error analysis is not adequately supported yet. Although some tools might provide the option to inspect every single annotation, custom transformation scripts are necessary to translate this information in a suitable format and, even then, the generated output (e.g. CSV or spreadsheet files with named entities and information on whether they have been classified as correct or incorrect) often lacks the context required for further analysis (e.g., the full text or sentence, the concordance of the entity, etc.).

Taking inspiration from the Open Data movement, we propose that NEL evaluations need to be more transparent in order to enhance both reproducibility and rapid development of new tools. In our opinion, **transparent benchmarking** systems need to fulfill the following six requirements:

- (i) *widely recognized metrics* - precision, recall, F1, accuracy or clustering measures.
- (ii) *explained evaluation runs* - we should not only be able to see the evaluation results, but also the classification into test results like false positives or false negatives or even into more fine-grained error classes if possible;
- (iii) *integrated visual analysis methods* - drill-down analysis should be used for inspecting and debugging the results;
- (iv) *support for resource versioning* - is needed in order to allow an evaluation to run with a previous version of a KB (e.g., run with DBpedia 3.9 or DBpedia 2015-10);
- (v) *reproducible settings for the annotator tools and the annotation tasks* - the settings that correspond to results published in a paper should be publicly available.
- (vi) *machine-readable annotation guidelines* - while annotation guidelines like those from TAC-KBP [11] are publicly available, it is hard to do reasoning with them or to combine them according to the task due to the fact that they are not available in a machine readable-format like RDF or its derivatives.

Since the current generation of annotation tools rarely publish their best settings and annotation guidelines are not really available in machine-readable formats to the best of our knowledge, the last two steps can be considered research topics onto themselves for now.

This paper presents *Orbis* [1], a system that addresses the first four of these requirements. Orbis is a versatile framework for performing NEL evaluations which supports standard metrics such as precision, recall and F1. The system visualizes gold standard and annotator results within their context (i.e. the annotated document) using color coding to aid experts in identifying correct and incorrect links. Orbis also provides links to the KBs against which an entity has been grounded, supporting easy lookup of entity properties within the target KB (see Figure 2). Modes for comparing multiple evaluations (i.e. the gold standard and two or more evaluations performed on it) and overview pages which highlight documents for which the evaluation results have been improved or deteriorated aid experts in quickly identifying strengths and weaknesses in their components and in addressing them. These comparative evaluations can be used to outline differences between systems, evaluation settings and various gold standard versions.

The remainder of this paper is organized as follows: Section 2 discusses related work on benchmarking and evaluating Named Entity Linking. Section 3 then provides a short overview of Orbis, its main components and capabilities. The use cases in Section 4 demonstrate how Orbis has been successfully applied to improving the performance of our own Named Entity Linking components, but also to spotting errors or addressing resource versioning issues, due to its advanced drill-down analysis capabilities. Section 5 presents a short outlook and conclusions.

2. Related Work

Due to the complexity of naming conventions there is no universally correct method for annotating entities. People names, for instance, can sometimes include titles (e.g., *Prince Charles*) and location entities can contain references to the region or state (e.g., *Dallas, TX*). In order to clarify how human experts annotated entities during the creation of a

[1] Orbis will be published on HTW Chur's GitHub account (https://github.com/htwchur) in Q4 2018.

gold standard, annotation guidelines need to be included when datasets are published. Some general annotation rules have been established quite early through the NERC guidelines provided by the MUC-7 (1997) and CoNLL 2003 challenges [16]. In recent times, the focus has shifted towards NEL and the most influential annotation guidelines have been those popularized by the TAC-KBP Challenges [11] for longer news media textual content, and NEEL Microposts Challenges [16] for shorter content like tweets.

Due to the widespread adoption of semantic KBs like DBpedia or Wikidata various scenarios have appeared for the evaluation of named entities: recognition and classification, linking, typed evaluations, and so on. Cornolti [3], therefore, defined new evaluation types based on the content of the annotation tasks. Six annotation tasks have been initially included in the BAT framework, an automated evaluation system that measures per-task performance: Disambiguate to Wikipedia (D2W), Annotate to Wikipedia (A2W), Scored-annotate to Wikipedia (Sa2W), Concepts to Wikipedia (C2W), Scored concepts to Wikipedia (Sc2W) and Ranked-concepts to Wikipedia (Rc2W). The proposed naming convention for the experiment types is relatively easy to understand and clearly explains each experiment type.

A sequel to Cornolti's work, GERBIL [19] is a large-scale evaluation system that is focused on comparing the output of different NEL systems. GERBIL has been used in many NEL challenges (e.g., ESWC's Open Knowledge Extraction Challenges, ISWC's Semantic Web Challenges) and includes more than ten datasets in its online version. GERBIL supports gold standards in the NIF format (Natural Language Processing Interchange Format), an RDF format designed to allow the sharing of both textual and annotation resources, and to ease the interplay between NLP tools [8]. The experiment types that are supported include: Entity Recognition, Disambiguate to KB (D2KB), Entity Typing, Concept to KB (C2KB), Annotate to KB (A2KB), Entity Recognition and Typing to KB (RT2KB), and two tasks for Open Knowledge Extraction (OKE Tasks) that were used in the OKE Challenges [14] for Information Extraction. While GERBIL is useful for comparing the results of NEL systems, it does not provide efficient tools for investigating individual results.

The scorer [5] used in recent TAC-KBP challenges [11] offers the option to inspect individual results and evaluation runs, but only provides a command-line interface. Besides aggregated metrics, it also provides output suitable for a primary analysis of the run that labels each annotator mention as *correct*, *incorrect*, *extra* (i.e. a named entity does not occur in the gold standard) or *missing*. Although the created output lacks context, such as the text surrounding a mention, this kind of analysis is already quite helpful in identifying and understanding NEL errors.

While not immediately apparent, NEL evaluations are still plagued by errors, even though the number of good scorers is on the rise. Issues can appear due to different guidelines or taxonomies used during the initial annotation of the gold standards, changes between KB versions, redirects, links in multiple languages or even due to the scoring components. A taxonomy of error classes collected from multiple annotators and gold standards based on the most likely location where the error was triggered is presented in Braşoveanu et al. [2], together with examples of the five discussed error classes: Knowledge Base (KB), Dataset (DS), Annotator (AN), NIL Clustering (NIL), and Scorer (SE). The proposed taxonomy can also help KB or evaluation systems maintainers to spot errors in their tools, which makes it ideal as a basis for rapid debugging. A tool from the GERBIL ecosystem, EAGLET [10], presents similar ideas, but focuses mostly on classifying several error types (e.g., redirects or missing annotations) found in gold standards.

Several recent tools decided to focus on visual error analysis. An early system that is focused on a primary visual error analysis aimed at improving a system's output is presented in Heinzerling and Strube [7].

3. Method

Orbis is an extendable evaluation framework written in Python 3.6 which offers multiple evaluation modes, resource versioning, parallel evaluation runs, dataset normalization, and drill-down analyses. These features were built with a flexible pipeline system designed to help configure, modify and extend evaluation processes. Orbis addresses the need for transparent benchmarking and visual inspection of the evaluation runs.

3.1. System architecture

Orbis is composed of two main components: The Orbis control component and the evaluation pipeline. Orbis controller handles the evaluation runs that are configured using YAML configuration files and executes them within the

Orbis pipeline. The configuration files specify the components needed for an evaluation run, such as the corpus with the gold standard annotations, the automated annotator service to be used, data normalization steps, scoring and metrics computation algorithms, as well as the display formats for the pipeline outputs (e.g., output files, visualizations). Orbis can process multiple configuration runs in serial or parallel order. The Orbis runner orchestrates the different pipelines for each evaluation run and handles the details related to their execution.

The Orbis evaluation pipeline consists of the following three stages: (i) acquisition and normalization of NEL annotations, (ii) assessment of these annotations, and (iii) visualizations and analytics.

The *acquisition and normalization of the NEL annotations stage* processes annotations extracted from the gold standard and annotations obtained from the systems under test. NEL tools like Recognyze or Spotlight [4] are integrated into Orbis during this stage via their public web services. A built-in caching framework (*results cache*) speeds up evaluations and reduces the load imposed on third-party web services. During this stage, additional data normalization steps such as rewriting of redirects which allows the correct handling of synonyms, and resource versioning which helps in keeping track of the link changes between different Knowledge Base versions can be performed to prepare the data for the evaluation. The rewriting of redirects is supported through the creation of mappings files, whereas link updating is performed by integrating lenses which contain all the links from a gold standard as they are reflected in a certain Knowledge Base version (see Section 4.3).

Multiple operations are executed in the *the assessment stage*. First the acquired annotations are scored based on the scoring rules defined in the configuration file yielding the confusion matrix. These scores are then used to calculate metrics such as micro and macro precision, recall, and the F1 measure. As the focus of our own research lies on the correct linking of surface forms to their corresponding resources in a target Knowledge Base, we currently do not consider NIL Clustering, although future research will extend the Orbis framework to support the evaluation of annotations that are not contained in the Knowledge Base.

The *visualization and analytics stage* converts the results into different formats to generate human or machine readable outputs for further processing and comparative analytics such as JSON formatted files or CSV tables, generating HTML views for drill-down analysis of single entity annotations, and for transmitting them to time-series database like Graphite[2] to track NEL performance over time.

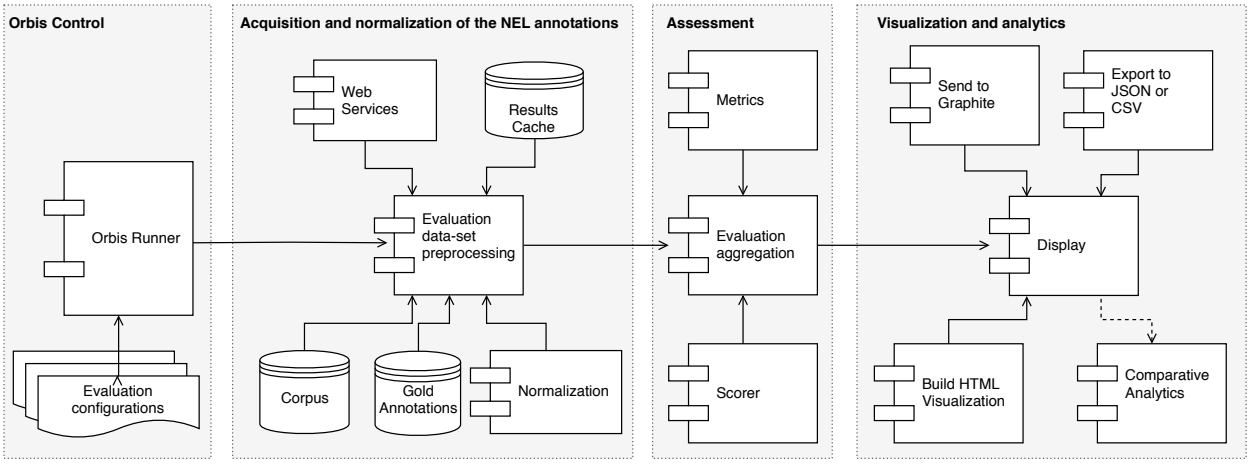

Fig. 1. Orbis system diagram

[2] https://graphiteapp.org/

3.2. Extensibility

Orbis uses a plugin system for all the stages from acquisition and normalization of resources, to assessment, and visualization and analytics. Plugins specified in the configuration files associated with the evaluation run are automatically recognized and integrated.

Currently, the plugin architecture supports both public and private web services endpoints, evaluation corpora, native (e.g., evaluations with or without overlaps) or third-party scorer plugins (e.g., TAC-KBP neleval [5]) and evaluation metrics. Scorer plugins define how to score the results. Metrics plugins compute measures such as precision, recall and F1-score. The output files or visualizations can also be controlled by defining new plugins.

Besides the plugin system, Orbis also supports an add-on system for applications that are useful in the context of using Orbis, but which are not part of an Orbis evaluation run. These applications are accessed by running the Orbis command line interface. Orbis detects installed add-ons automatically and displays them in a selection menu from where they can be executed. Add-ons allow data preprocessing as well as post processing and preparation of evaluation runs. Since Orbis requires its own corpus format, an add-on can be used to create this format from standards such as TAC-KBP neleval [5] or NIF [8]. Visualization of results from multiple evaluation runs can also be created by using an add-on, and there is even an add-on that automatically generates YAML configuration files for comparative evaluations covering combinations of different corpora, web services, lenses, mappings, filters and output formats.

4. Drill-down analysis with Orbis

Orbis has been created to allow drill-down analysis of NEL results. To improve transparency, a visualization of the results was added to the interface. This allows NEL tool developers to visually debug their systems and to compare NEL results with both gold standard annotations and the output obtained from other tools.

Orbis visually displays gold annotations and the result of the NEL process in text tabs next to each. Since linked entities are indicated using color coding, experts can quickly compare the named entity links annotated in the gold standard to the annotator results. These visualizations facilitate the detection of NEL errors in the annotator results, as well as in the gold standard annotations (Section 4.1). Orbis generates these visualizations using HTML to allow the results of each test document to be viewed and navigated using a web browser. An Orbis add-on based on the before mentioned visualization allows multiple annotator results to be displayed alongside the gold standard annotation, facilitating drill-down analysis that compare different evaluation settings and systems.

Most evaluation tools focus on aggregated performance metrics such as precision, recall, and F1. They, therefore, provide useful means for comparing systems and evaluation runs to each other but often also act as black boxes when it comes to understanding the strengths and weaknesses of NEL systems.

NEL researchers and system designers, in contrast, require potent means for performing drill-down analysis that help in *explaining* rather than only spotting changes to a system's performance, allow assessing gold standard quality, and classifying errors into annotation and dataset errors [2] that help in understanding and addressing shortcomings of their NEL systems and of the used evaluation datasets.

Our interface was designed based on the grammar of graphic principles outlined by Wilkinson [23], as well as on the two very influential taxonomies of Shneiderman [18] and Heer [6]. In general we try to follow the basic phases of visual analytics as expressed by Shneiderman: *overview* first, *zoom*, *filter*, then provide the *details-on-demand*, *relations (relate principle)*, *history* and *extract* [18]. Our *overviews* contain general corpus results (e.g., precision, recall, f1), then we can *zoom* into single documents and if needed we can use lenses to *filter* based on various criteria (Section 4.3) or provide *details-on-demand* about individual entities. A *history* can be kept in a time-series database (e.g., Graphite) or locally as CSV or JSON file objects to be used in the post analysis phase. The *relations (relate principle)* between various evaluation runs can easily be explored via interface since it allows to compare runs from the same or multiple annotators. This relational analysis can typically be considered the start of the post analysis phase. The *extract* phase is implemented in various parts of the interface via highlighting (e.g., for example in the overviews table), where all the problematic documents are highlighted. Using these principles helped us enhance the drill-down analysis features as described in the following use cases.

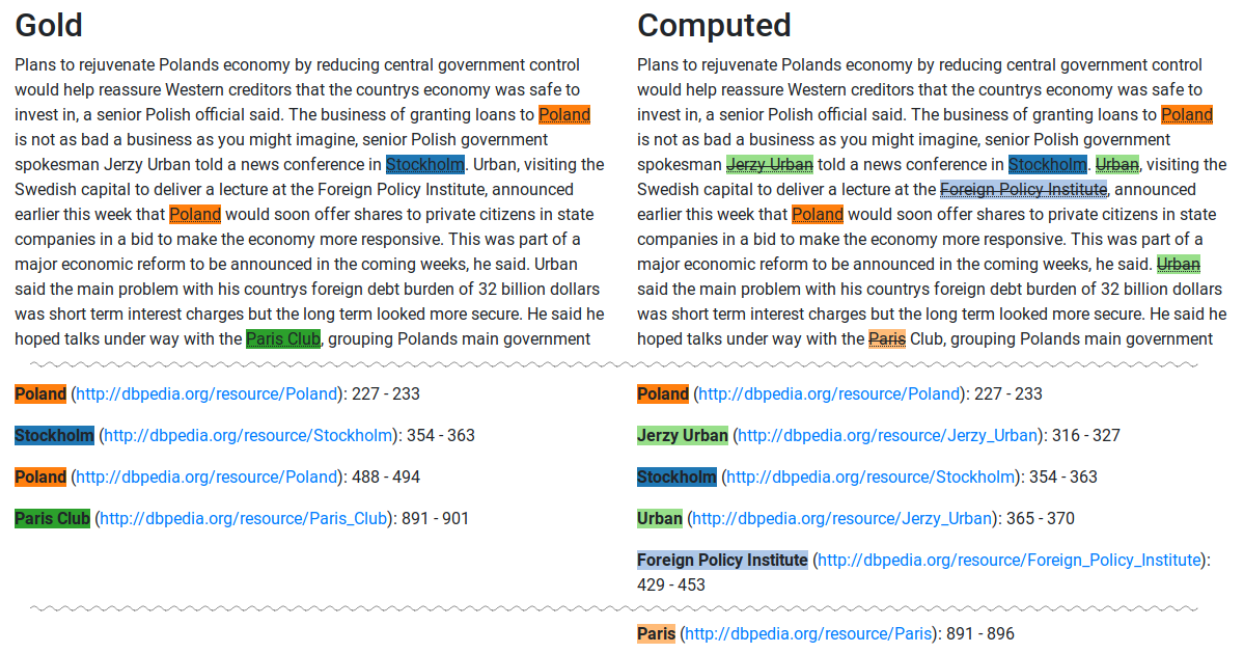

Fig. 2. A cropped screenshot of the results as generated by Orbis demonstrating the results of file 106 of the Reuters128 evaluation corpus as annotated with our NEL component. Left demonstrates the gold standard, right demonstrates the results returned by the annotator system. The upper half highlights the annotations in the used test document, while the lower half lists the annotations in textual order. Matching colors indicate identical resources.

4.1. Improving NEL tools

Focusing on the improvement of NEL tools, Orbis does not only determine if a change helped improving aggregated NEL performance metrics, but also enables the developer to perform drill-down analysis that helps in identifying error patterns.

Figure 2 shows a cropped screenshot of the output of the Orbis framework consisting of two columns, left for the gold standard of the evaluation corpus, right for the evaluated NEL annotator tool. The upper part highlights surface forms in the test document, while the lower part lists all surface forms chronologically. The colored highlighting of the entities' surface forms demonstrates (in)consistencies between the corpus' gold standard and the annotator, whereas identical colors indicate correctness (for example the annotation *Poland*), while a different coloring for the same surface form points to an incorrect annotation (*Paris [Club]*). A surface form only highlighted in the left column generally indicates missing annotations, while surface forms only highlighted in the right column generally indicate extra annotations (*Foreign Policy Institute*). As mentioned in Section 3.1 our focus lies on entities with a link to a Knowledge Base, NIL entities being currently ignored on both sides.

While the visualization of named entities within the analyzed document makes spotting problems and error patterns fast and obvious, it is the list of annotations in the lower part of the visualized output that really helps in understanding and classifying error types and guides NEL designers in optimizing their efforts towards the most effective improvements of their systems. Besides the surface form itself, also the link to the resource in the target Knowledge Base, as well as the string index in the text itself is listed. Examining the elements of this list, together with their corresponding resource in the KB helps in detecting

- incorrectly labelled resources in the KB, such as the resource of *Fritz Kraatz* which contains the label *Barack Obama* [20].
- incorrect attributes in the KB, for example a person that is also an organization [22];

- problems with the preprocessing processes like the ambiguous extraction of *Home* from *Home, Kansas*;
- mistakes in the disambiguation processes, as for example an incorrect mapping to the wrestler *Ace Steel* who uses the alter ego *Donald Trump*.

Not illustrated in the cropped screenshot of Figure 2 is an area located at the top of the tool containing the general performance indicators (e.g., precision, recall, F1), as well as the number of true positives, false positives and false negatives in the observed file. Comparing these numbers between different runs helps fast detecting significant differences between two or more evaluation runs which can then be compared to examine the impact of a change in the NEL algorithm.

4.2. Improving corpus quality

Most of the traditional evaluation tools (including GERBIL) are focused on delivering the evaluation results. After the Open Data movement took off, a new trend has emerged: transparent evaluations, or alternatively open evaluation systems. Orbis allows system developers to examine both the gold standards and the system results side by side, therefore, it can be considered an open evaluation system.

Drill-down analyses on the level of individual annotations lead to a thorough examination of each test document and each annotation returned by the evaluated system. Generally speaking, an expert tries to answer the question of why the evaluated system thinks that the returned annotation is correct to determine systematic problems with the annotator and ways to mitigate these. But - in very seldom cases - the evaluated annotators may also reveal annotations that are correct, but not contained in the used gold standard.

Such an example is represented by *Jerzy Urban* as it can easily be seen in Figure 2. The gold standard indicates that Urban, a Polish politician, does does not have a clear link to a KB entry and the entity has, therefore, been marked as NIL. Nevertheless, the evaluated NEL tool has successfully grounded the entity to the target Knowledge Base version used during the creation of the gold standard (DBpedia 3.9). Although it is displayed by Orbis as an extra annotation, and also treated as such, it is in fact not a problem with the annotator but with the gold standard itself. Possible reasons for this flaw within the gold standard are that (i) either the referenced resource was not contained in the KB at the time of the corpus creation; or (ii) the creators of the gold standard were not able to identify the correct resource or a link for it in the respective Knowledge Base version. Although we do not propose a workflow to fix or dismiss such issues, Orbis can help in spotting them and, therefore, improves transparency.

4.3. Resource versioning

Two large error classes discussed in [2] focused on the issue of tracking changes to the links that correspond to a mention: KB errors and NIL Clustering errors. Through the implementation of lenses, Orbis provides means for easily addressing these error classes.

The basic idea behind lenses, as expressed in Figure 3, is that new annotation sets containing the various entity links as they are represented in a certain KB version (e.g., DBpedia 3.9, DBpedia 2015-04, DBpedia 2016-04, etc.) are generated manually or automatically from the existing gold standard annotations (see [1]). While an argument against using such additional annotation sets can easily be made by stating that there should be no changes done to an existing gold standard, in reality such changes would serve a dual-purpose: (i) they would help to reproduce results on older versions of KBs; and (ii) would enable automatic updating gold standards to later version of KBs. We have mainly taken the first approach and used lenses in order to reproduce old results. We have used full builds of the respective KBs whenever possible. When the builds are too large, there is always the possibility to slice the builds via a dedicated slicer component [12] or custom code and include only the entity types used in the respective evaluation. Using lenses in order to update gold standards generally requires the permission of the original creators of the datasets, especially if they were published under a restrictive license.

Perhaps another important use case for lenses is to make sure that NEL tools remain competitive in evaluations and production environments regardless of the KB version used. This is especially important if the tools in question also have commercial settings, where usually the latest KB versions need to be deployed.

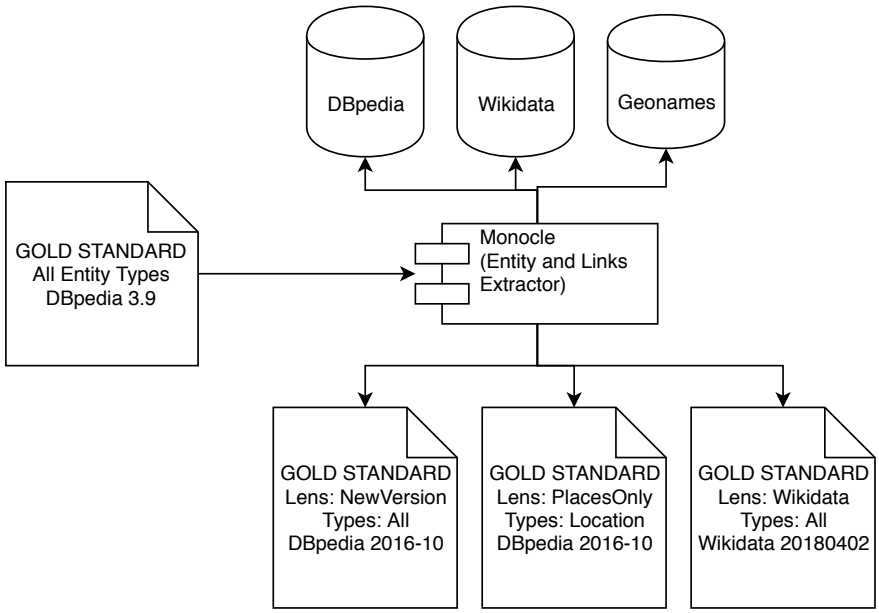

Fig. 3. Creation of lenses that help with resource versioning in Orbis.

4.4. Evaluation Types and Datasets

Orbis currently supports the following evaluation types: *Entity Recognition (ER)*, *Disambiguate to KB (D2KB)*, *Entity Typing (ET)* and *Lenses Evaluation (LENS)*.

Several annotator tools have already been integrated: Recognyze (an updated version of [21]), Spotlight [4], Babelnet [13] and AIDA [9]. Not all annotators publish their best settings necessarily, some of them advising users to experiment until they find the best settings for their experiments, therefore where such best settings have not been published those annotators were integrated using their most recently published settings. All these tools currently return DBpedia results. Multiple tools will be added in the near future.

Several datasets have already been integrated into Orbis, each dataset being used in at least one if not multiple evaluation types, depending on the number of annotation sets available for it. *Reuters128* is part of the larger *N3 collection* [17] and contains texts with popular entities extracted from the classic Reuters corpora. *OKE2015* [14] and *OKE2016* [15] are two datasets used during the SemEval at ESWC conferences which contain short biographic sentences selected from Wikipedia/DBpedia abstracts. These datasets can be used with the first three evaluation types without any changes to the core Orbis infrastructure.

Evaluations with lenses (e.g., different Knowledge Base version, relations) are supported through the integration of the StoryLens dataset [1], but in some cases a small plugin might be needed if the rules for a particular lens are not already included in the list of available Orbis evaluations.

5. Outlook and Conclusions

This paper described Orbis, an extensible framework for performing Named Entity Linking evaluations that provides visualizations of evaluation results which enable researchers and system architects to quickly inspect errors. In contrast to other NEL evaluation tools which often only provide aggregated metrics or very rudimentary information on linking errors, Orbis displays gold standard and NEL output within their textual context, provides information on all linked entities and means to obtain further background information on these entities.

The visualizations allow researchers to quickly compare the performance of two systems with each other and the gold standard. This evaluation mode is particularly useful in assessing the effects of architectural changes and in evaluating the strengths and weaknesses of different NEL systems.

Since Orbis is a flexible framework and offers an affordable option for building new evaluation use cases, it can easily be argued that it is in fact a framework designed to help build evaluation infrastructure.

As outlined through the presented use cases, Orbis significantly lowers the effort required to perform drill-down analysis which in turn enable researchers to locate a problem in algorithms, machine learning components, gold standards and data sources more quickly, leading to a more efficient allocation of research efforts and developer resources.

Future work will be focused on (i) integrating statistical significance tests such as the Wilcoxon Rank Sum test into the Orbis platform; (ii) creating plugins for tracking and publishing evaluation results; (iii) developing support for additional evaluation types such as Concepts to KB and sentiment analysis; and (iv) integrating more datasets and tools for each task.

Acknowledgements

The research presented in this paper has been conducted as part of the DISCOVER Project (www.htwchur.ch/discover), funded by the Swiss Commission for Technology and Innovation (CTI). Adrian Braşoveanu's work was also supported through the InVID project funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 687786.

References

[1] A. M. P. Braşoveanu, L. J. Nixon, and A. Weichselbraun. Storylens: A multiple views corpus for location and event detection. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*, Novi Sad, Serbia, 2018. ACM. doi: 10.1145/3227609.3227674. URL http://doi.acm.org/10.1145/3227609.3227674.

[2] A. M. P. Braşoveanu, G. Rizzo, P. Kuntschick, A. Weichselbraun, and L. J. Nixon. Framing named entity linking error types. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 266–271, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html.

[3] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In D. Schwabe, V. A. F. Almeida, H. Glaser, R. A. Baeza-Yates, and S. B. Moon, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 249–260. International World Wide Web Conferences Steering Committee / ACM, 2013. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488411. URL http://dl.acm.org/citation.cfm?id=2488411.

[4] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity extraction. In M. Sabou, E. Blomqvist, T. D. Noia, H. Sack, and T. Pellegrini, editors, *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124. ACM, 2013. doi: 10.1145/2506182.2506198. URL http://dl.acm.org/citation.cfm?id=2506182.

[5] B. Hachey, J. Nothman, and W. Radford. Cheap and easy entity evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 464–469. The Association for Computer Linguistics, 2014. ISBN 978-1-937284-73-2. doi: 10.3115/v1/P14-2076. URL http://aclweb.org/anthology/P/P14/P14-2076.pdf.

[6] J. Heer and B. Shneiderman. Interactive Dynamics for Visual Analysis. *Communications of the ACM*, 55:45–54, 2012. doi: 10.1145/2133416.2146416. URL https://dl.acm.org/citation.cfm?id=2133821.

[7] B. Heinzerling and M. Strube. Visual error analysis for entity linking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, System Demonstrations*, pages 37–42. ACL, 2015. ISBN 978-1-941643-99-0. doi: 10.3115/v1/P15-4007. URL http://aclweb.org/anthology/P/P15/P15-4007.pdf.

[8] S. Hellmann, J. Lehmann, S. Auer, and M. Nitzschke. NIF combinator: Combining NLP tool output. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, editors, *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, volume 7603 of *Lecture Notes in Computer Science*, pages 446–449. Springer, 2012. ISBN 978-3-642-33875-5. doi: 10.1007/978-3-642-33876-2_44. URL http://dx.doi.org/10.1007/978-3-642-33876-2_44.

[9] 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 *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792. ACL, 2011. ISBN 978-1-937284-11-4. doi: 10.3115/v1/D11-1072. URL http://www.aclweb.org/anthology/D11-1072.

[10] K. Jha, M. Röder, and A. N. Ngomo. All that glitters is not gold - rule-based curation of reference datasets for named entity recognition and entity linking. In E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, editors, *The Semantic Web - 14th*

International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I, volume 10249 of *Lecture Notes in Computer Science*, pages 305–320, 2017. ISBN 978-3-319-58067-8. doi: 10.1007/978-3-319-58068-5_19. URL https://doi.org/10.1007/978-3-319-58068-5_19.

[11] H. Ji and J. Nothman. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end kbp. In *Eighth Text Analysis Conference (TAC)*. NIST, 2016. URL https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf.

[12] E. Marx, S. Shekarpour, T. Soru, A. M. P. Braşoveanu, M. Saleem, C. Baron, A. Weichselbraun, J. Lehmann, A. N. Ngomo, and S. Auer. Torpedo: Improving the state-of-the-art RDF dataset slicing. In *11th IEEE International Conference on Semantic Computing, ICSC 2017, San Diego, CA, USA, January 30 - February 1, 2017*, pages 149–156, San Diego, CA, USA, 2017. IEEE Computer Society. ISBN 978-1-5090-4284-5. doi: 10.1109/ICSC.2017.79. URL https://doi.org/10.1109/ICSC.2017.79.

[13] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/291.

[14] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, D. Garigliotti, and R. Navigli. Open knowledge extraction challenge. In F. Gandon, E. Cabrio, M. Stankovic, and A. Zimmermann, editors, *Semantic Web Evaluation Challenges - Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, volume 548 of *Communications in Computer and Information Science*, pages 3–15, Berlin, Germany, 2015. Springer. ISBN 978-3-319-25517-0. doi: 10.1007/978-3-319-25518-7_1. URL https://doi.org/10.1007/978-3-319-25518-7_1.

[15] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, R. Meusel, and H. Paulheim. The second open knowledge extraction challenge. In H. Sack, S. Dietze, A. Tordai, and C. Lange, editors, *Semantic Web Challenges - Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*, volume 641 of *Communications in Computer and Information Science*, pages 3–16, Berlin, Germany, 2016. Springer. ISBN 978-3-319-46564-7. doi: 10.1007/978-3-319-46565-4_1. URL https://doi.org/10.1007/978-3-319-46565-4_1.

[16] G. Rizzo, B. Pereira, A. Varga, M. van Erp, and A. E. C. Basave. Lessons learnt from the named entity recognition and linking (NEEL) challenge series. *Semantic Web*, 8(5):667–700, 2017. doi: 10.3233/SW-170276. URL https://doi.org/10.3233/SW-170276.

[17] M. Röder, R. Usbeck, S. Hellmann, D. Gerber, and A. Both. N^3 - A collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 3529–3533, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/856.html.

[18] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages, IEEE VL 1996*, pages 336–343. IEEE, 1996. doi: 10.1109/MSP.2014.80. URL https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf.

[19] 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 *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, pages 1133–1143, 2015. doi: 10.1145/2736277.2741626. URL http://doi.acm.org/10.1145/2736277.2741626.

[20] A. Weichselbraun and P. Kuntschik. Mitigating linked data quality issues in knowledge-intense information extraction methods. In *WIMS 2017*, pages 1–12. ACM Press, 2017. ISBN 978-1-4503-5225-3. doi: 10.1145/3102254.3102272. URL http://dl.acm.org/citation.cfm?doid=3102254.3102272.

[21] A. Weichselbraun, D. Streiff, and A. Scharl. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools*, 24(2):1–31, 2015. doi: 10.1142/S0218213015400084. URL https://doi.org/10.1142/S0218213015400084.

[22] A. Weichselbraun, P. Kuntschik, and A. M. P. Braşoveanu. Mining and leveraging background knowledge for improving named entity linking. In *Proceedins of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018)*, Novi Sad, Serbia, 2018. ACM. doi: 10.1145/3227609.3227670. URL http://doi.acm.org/10.1145/3227609.3227670.

[23] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Statistics and Computing. Springer-Verlag New York, Secaucus, NJ, USA, 2005. ISBN 0387245448. doi: 10.1007/0-387-28695-0. URL https://www.springer.com/de/book/9780387245447.