Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job

Marieke van Erp[◊], **Pablo N. Mendes**[*], **Heiko Paulheim**[∓], **Filip Ilievski**[◊], **Julien Plu**[°], **Giuseppe Rizzo**[‡], **Jörg Waitelonis**[∞]

[◊]Vrije Universiteit Amsterdam, Netherlands {marieke.van.erp,f.ilievski}@vu.nl

[*]IBM Research Almaden, USA pnmendes@us.ibm.com

[∓]Data and Web Science Group, University of Mannheim, Germany heiko@informatik.uni-mannheim.de

[°]EURECOM, France plu@eurecom.fr

[‡]ISMB, Italy giuseppe.rizzo@ismb.it

[∞]Hasso-Plattner-Institute Potsdam, Germany joerg.waitelonis@hpi.de

Abstract

Entity linking has become a popular task in both natural language processing and semantic web communities. However, we find that the benchmark datasets for entity linking tasks do not accurately evaluate entity linking systems. In this paper, we aim to chart the strengths and weaknesses of current benchmark datasets and sketch a roadmap for the community to devise better benchmark datasets.

Keywords: entity linking, evaluation, benchmark

1. Introduction

Since the 90's, recognizing and linking entities has been a popular research topic. Initially, research attempts focused on identifying and classifying atomic information units in text (Named Entity Recognition and Classification). Later on, research into linking mentions to external resources of knowledge bases referents (Named Entity Linking) was introduced. This triggered numerous research initiatives such as CoNLL (Tjong Kim Sang and De Meulder, 2003), ACE (Doddington et al., 2004), TAC-KBP (McNamee, 2009), NEEL (Cano et al., 2014; Rizzo et al., 2015), SemEval (Moro and Navigli, 2015), or ERD (Carmel et al., 2014), aiming at building common and general benchmark datasets to test, adapt, and improve entity recognition and linking.

Benchmark datasets have often been treated as black boxes by published research, making it difficult to interpret efficacy improvements in terms of individual contributions of algorithms and/or labeled data. This scattered landscape of datasets and measures leads to a misleading interpretability of the experimental results, which makes the performance evaluation of novel approaches against the state of the art rather difficult and open to several interpretations and questions.

In this paper, we aim to highlight the strengths and weaknesses of common benchmark datasets developed for evaluation of entity linking systems. We inspect heterogeneous benchmark datasets in terms of genre such as newswire, blog posts, and microblog posts. Dataset characteristics have been previously reviewed in (Ling et al., 2015) as part of the overall analysis of the entity linking task, in (Steinmetz et al., 2013) with focus on candidate generation, and in GERBIL (Usbeck et al., 2015). Our work is complementary to all, as we focus on deepening the analysis of the benchmark datasets. We approach this by highlighting strengths and weaknesses of these datasets through quantifiable aspects such as entity overlap, dominance, and popularity. We hope that this work may foster metric-aware and less biased benchmark datasets to be created, as well as it may stimulate a more sensitive discussion of results produced using those benchmarks.

The paper is organized as follows: Section 2. gives an overview of the benchmark datasets we investigate. We detail their characteristics in Section 3. In Section 4. we report insights and key interpretations, and we devise a roadmap of current and future work. Section 5. concludes this paper.

2. Datasets

This section presents the datasets that we analysed. The datasets are listed in alphabetical order and we describe their main characteristics.

AIDA-YAGO2 Dataset

The AIDA-YAGO2 dataset (Hoffart et al., 2011)[1] is an extension of the CoNLL 2003 entity recognition task dataset (Tjong Kim Sang and De Meulder, 2003). It is based on news articles published between August 1996 and August 1997 by Reuters. Each entity is identified by its YAGO2 entity name, Wikipedia URL, and, if available, by Freebase Machine ID.

2014 / 2015 NEEL

The 2014 Microposts dataset (Cano et al., 2014)[2] consists of 3,504 tweets extracted from a much larger collection of over 18 million tweets. The tweets were provided by the Redites project, which covers event-annotated tweets collected for a period of 31 days between July 15th, 2011 and August 15th, 2011. It includes multiple noteworthy events,

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

[2] http://scc-research.lancaster.ac.uk/workshops/microposts2014/challenge/index.html

Corpus	Type	Domain	Doc. Length	Format	Encoding	License
AIDA-YAGO2	news	general	medium	TSV	ASCII	Source data via agreement
2014/2015 NEEL	tweets	general	short	TSV	ASCII	Open
OKE2015	encyclopedia	general	long	NIF/RDF	UTF8	Open
RSS-500	news	general	medium	NIF/RDF	UTF8	Open
WES2015	blog	science	long	NIF/RDF	UTF8	Open
WikiNews	news	general	medium	stand-off XML	UTF8	Open

Table 1: General characteristics for analysed datasets

such as the death of Amy Winehouse, the London Riots, and the Oslo bombing.
The 2014 Microposts challenge dataset was created to benchmark automatic extraction and linking entities. The corpus is split into a train and a test set.
The 2015 corpus (Rizzo et al., 2015)[3] contains more tweets (6,025) and covers more noteworthy events from 2011 and 2013 (e.g. the Westgate Shopping Mall shootout), as well as tweets extracted from the Twitter firehose in 2014. The training set is built on top of the entire corpus of the #Microposts2014 NEEL challenge. It was further extended to include entity types and NIL references.

OKE2015

The Open Knowledge Extraction Challenge 2015 (OKE2015) (AndreaGiovanniNuzzolese et al., 2015)[4] corpus consists of 197 sentences from Wikipedia articles. The annotation task focused on recognition (including co-reference), classification according to the Dolce Ultra Lite classes,[5] and linking of named entities to DBpedia. The annotation set was created using the Wikipedia links found in the articles, extended with automatic anaphora resolution, and detection of emerging entities. The annotation set was manually reviewed and fixed. The corpus was split into a train and test set containing a similar number of sentences: 96 in the training set, and 101 in the test set.

RSS-500-NIF-NER

The RSS-500 dataset (Röder et al., 2014)[6] contains data from 1,457 RSS feeds, including major international newspapers. The dataset covers many topics such as business, science, and world news. The initial 76-hour crawl resulted in a corpus which contained 11.7 million sentences. Out of these, 500 sentences were manually chosen to be included in the RSS500 corpus. This set of sentences was annotated by one researcher. The chosen sentences contain a formal relation (e.g. "..who was born in.." for dbo:birthPlace), that should occur more than 5 times in the 1% corpus. This corpus was used for evaluation purposes in (Gerber et al., 2013).

WES2015

The WES2015 dataset was originally created to benchmark information retrieval systems (Waitelonis et al., 2015).[7] It contains 331 documents annotated with DBpedia entities. The documents originate from a blog about history of science, technology, and art.[8] The dataset also includes 35 annotated queries inspired by the blog's query logs, and relevance assessments between queries and documents.
The WES2015 dataset is available as NIF2 dump[9], as well as in RDFa (Adida et al., 2012) format annotated within the HTML source of the blog articles.

WikiNews/MEANTIME

The WikiNews/MEANTIME (hereafter referred to as 'Wikinews') dataset is a benchmark dataset that was compiled by the NewsReader project (Minard et al., 2016).[10] This corpus consists of 120 Wikinews articles, grouped in four sub-corpora: Airbus, Apple, General Motors and Stock Market. These are annotated with entities in text, including links to DBpedia, events, temporal expressions and semantic roles. This subset of WikiNews news articles was specifically selected to represent domain entities and events from the financial aspect of the automotive industry.

3. Dataset Characteristics

In this section, we describe the analyses we perform on the benchmark datasets.

3.1. Document Type

The documents which comprise benchmarking datasets can vary along several dimensions:

- **Type of discourse:** news articles, tweets, transcriptions, blog articles, scientific articles, government/medical reports
- **Topical domain:** science, sports, politics, music, catastrophic events, general (cross-domain)
- **Document length (in terms of number of tokens)**: long, medium, short
- **Format:** document format (TSV, CoNLL, NIF), stand-off vs. in inline annotation

[3] http://scc-research.lancaster.ac.uk/workshops/microposts2015/challenge/index.html
[4] https://github.com/anuzzolese/oke-challenge
[5] http://stlab.istc.cnr.it/stlab/WikipediaOntology/
[6] https://github.com/AKSW/n3-collection
[7] http://yovisto.com/labs/wes2015/wes2015-dataset-nif.rdf
[8] http://blog.yovisto.com/
[9] http://s16a.org/node/14
[10] http://www.newsreader-project.eu/results/data/wikinews

- **Character encoding:** Unicode, ASCII, URL-encoding
- **Licensing**: open, closed

Table 1 summarizes the document type characteristics of the corpora we analyze, already exposing notable diversity among the datasets with respect to the considered set of aspects.

3.2. Entity, surface form and mention characterization

In this section, we analyze and compare the coverage of entities and entity mentions in the different datasets along three dimensions: entity overlap, entity distribution, and entity types.

Entity Overlap In Table 2, we present the entity overlap between the different benchmark datasets. Each row in the table represents the percentage of unique entities present in that dataset that are also represented in the other datasets. As the table illustrates, there is a fair amount of overlap between the entities in the Wikinews dataset and the other benchmark datasets. The overlap between the NEEL2014 and NEEL2015 datasets is explained by the fact that the latter is an extension of the former. The WES2015 dataset has the least in common with the other datasets.

Confusability Let the true confusability of a surface form s be the number of meanings that this surface form can have. As new places, organizations and people are named every day, without access to an exhaustive collection of all named entities in the world, the true confusability of a surface form is unknown. However, we can estimate the confusability of a surface form through the function $A(s) : S \rightarrow \mathbb{N}$ that maps a surface form to an estimate of the size of its candidate mapping, such that $A(s) = |C(s)|$. The confusability of, for example, a place name offers only a rough *a priori* estimate of how difficult it may be to disambiguate that surface form. Observation of annotated occurrences of this surface form in a text collection allows us to make more informed estimates. We show the average number of meanings denoted by a surface form, indicating the confusability, as well as complementary statistical measures on the datasets in Table 3. In this table, we observe that most datasets have a low number of average meanings per surface form, but there is a fair amount of variation, i.e. number of surface forms that can refer to a meaning. In particular, the OKE2015 and Wikinews/MEANTIME datasets stand out in their high number of maximum meanings per surface form and standard deviations.

Given a surface form, some senses are much more *dominant* than others – e.g. for the name 'berlin', the resource `dbpedia:Berlin` (Germany) is much more 'talked about' than `Berlin,_New_Hampshire` (USA). Therefore, we also take into account estimates of *prominence* and *dominance* as:

Prominence Let the true prominence of a resource r_i be the percentage of other resources $r_k \in R$, which are less known than r_i. Let the prominence estimate $Pr(r_i)$ be the relative frequency with which the resource r_i appears linked on Wikipedia compared to the frequency of all other resources in R. Formally:

$$Pr(r_i) = \frac{\sum_{s \in S} |WikiLinks(s, r_i)|}{\sum_{s \in S, r \in R} |WikiLinks(s, r)|}$$

We estimate entity prominence through PageRank (Page et al., 1999). Some entities which are linked from only a few, but very prominent entities are also considered prominent. *Goethe's Faust*, for example, only has a few links, but one of those is *Goethe*, which is considered a prominent entity, and thus, *Goethe's Faust* would also be prominent.

Figure 1 depicts the PageRank distribution of the DBpedia based benchmarks compared to each other, as well as compared to the overall PageRank distribution in DBpedia.[11] The figure illustrates that all investigated benchmarks favor entities that are much more popular than average entities in DBpedia. Thus, the benchmarks show a considerable bias towards head entities. However, the whiskers of the box plots also show that all benchmarks contain long tail entities (i.e., all benchmarks contain some entities with minimum PageRank), and almost all of them also contain the DBpedia entity with the highest PageRank value (i.e., United States).

Evaluating against a corpus with a tendency to focus strongly on prominent entities may cause some issues. Entity Linking systems that include the global popularity of entities in their approach can reach very good results (Tristram et al., 2015), but these can hardly be transferred to other settings.

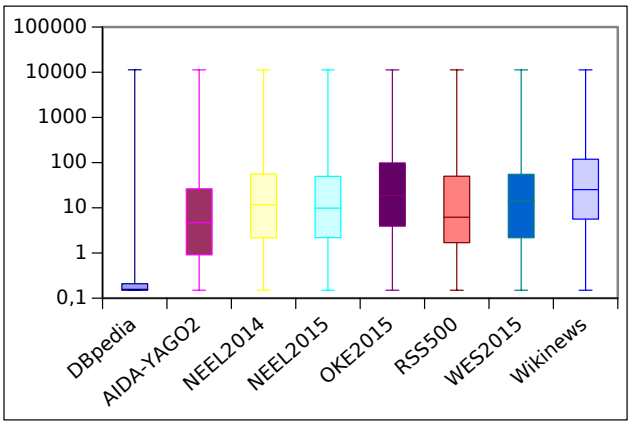


Figure 1: Distribution of DBpedia entity PageRank in the analyzed benchmarks. The leftmost bar shows the overall PageRank distribution in DBpedia as a comparison. The boxes depict the PageRank of the 25% of the instances with a PageRank above and below the median, respectively, while the whiskers capture the full distribution.

Dominance Let the true dominance of a resource r_i for a given surface form s_i be a measure of how commonly r_i is meant with regard to other possible meanings when s_i is used in a sentence. Let the dominance estimate $D(r_i, s_i)$ be the relative frequency with which the resource r_i appears in

[11] We use the DBpedia PageRank from `http://people.aifb.kit.edu/ath/`.

	AIDA-YAGO2	NEEL2014	NEEL2015	OKE2015	RSS500	WES2015	Wikinews
AIDA-YAGO2 (5596)	-	327 (5.87)	451 (8.06)	0 (0)	70 (1.26)	269 (4.8)	65 (1.16)
NEEL2014 (2380)	327 (13.73)	-	1630 (68.49)	57 (2.39)	61 (2.56)	294 (12.35)	67 (2.82)
NEEL2015 (2800)	451 (16.11)	1630 (58.21)	-	56 (2)	71 (2.54)	222 (7.93)	72 (2.57)
OKE2015(531)	0 (0)	57 (10.73)	56 (10.55)	-	13 (2.44)	149 (28.06)	21 (3.95)
RSS500 (849)	70 (8.24)	61 (7.18)	71 (8.36)	13 (1.53)	-	27 (3.18)	16 (1.88)
WES2015 (7309)	269 (3.68)	294 (4.02)	222 (3.04)	149 (2.04)	27 (0.16)	-	48 (0.66)
Wikinews (279)	65 (23.30)	67 (24.01)	72 (25.81)	21 (7.53)	16 (5.73)	48 (17.20)	-

Table 2: Entity overlap in the analyzed benchmark datasets. Behind the dataset name in each row the number of unique entities present in that dataset is given. For each datasets pair the overlap is given in number of entities and percentage (in parentheses).

Corpus	Average	Min.	Max.	σ
AIDA-YAGO2	1.08	1	13	0.37
2014 NEEL	1.02	1	3	0.16
2015 NEEL	1.05	1	4	0.25
OKE2015	1.11	1	25	1.22
RSS-500	1.02	1	3	0.16
WES2015	1.06	1	6	0.30
Wikinews	1.09	1	29	1.03

Table 3: Confusability stats for analyzed datasets. Average stands for average number of meanings per surface form, Min. and Max. stand for the minimum and maximum number of meanings per surface form found in the corpus respectively, and σ denotes the standard deviation.

Wikipedia links where s_i appears as the anchor text. Formally:

$$D(r_i, s_i) = \frac{|WikiLinks(s_i, r_i)|}{\forall_{r \in R}|WikiLinks(s_i, r)|}$$

The dominance statistics for the analysed datasets are presented in Table 4. The dominance scores for all corpora are quite high and the standard deviation is low, meaning that in vast majority of cases, a single resource is associated with a certain surface form in the annotations, creating a low of variance for an automatic disambiguation system.More statistics for dominance can be found on the GitHub page.

Corpus	Dominance	Max	Min	σ
AIDA-YAGO2	0.98	452	1	0.08
2014 NEEL	0.99	47	1	0.06
2015 NEEL	0.98	88	1	0.09
OKE2015	0.98	1	1	0.11
RSS-500	0.99	1	1	0.07
WES2015	0.97	1	1	0.12
Wikinews	0.99	72	1	0.09

Table 4: Dominance stats for analyzed datasets.

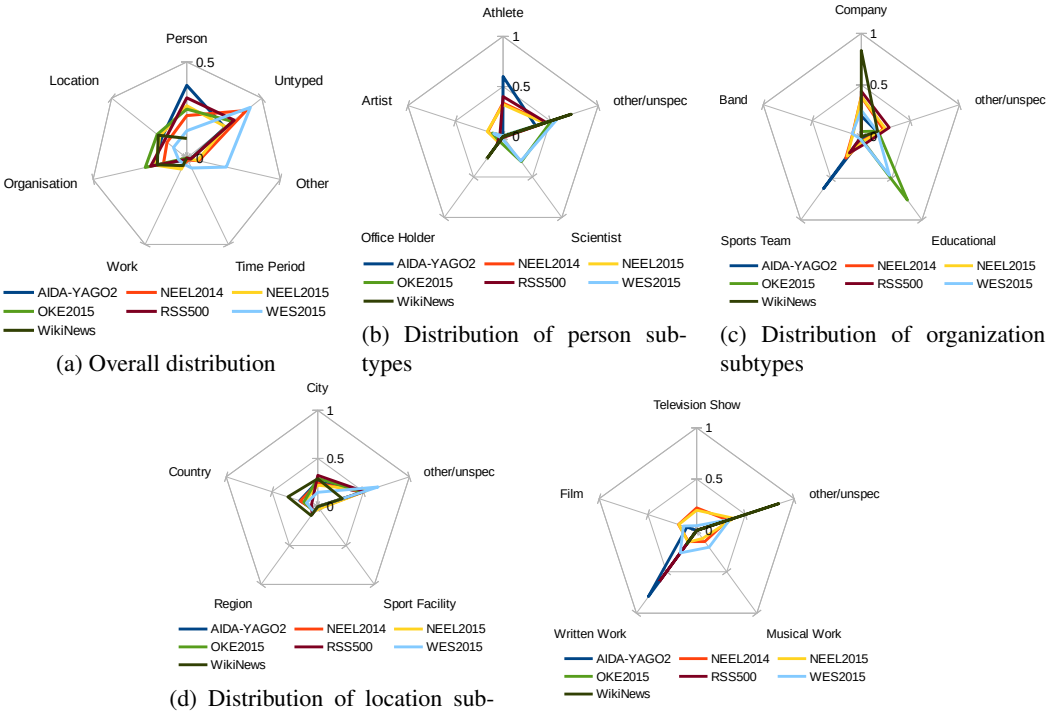
Corpora that contain resources with high confusability, low dominance and low prominence are considered more difficult to disambiguate. This is due to the fact that such corpora require a more careful examination of the context of each mention before algorithms can choose the most likely disambiguation. In cases with low confusability, high prominence and high dominance, simple popularity-based baselines that ignore the context of the mention can already perform quite accurately.

Entity types Entities characterized with certain semantic types may be more difficult to disambiguate than others. For example, while country and company names (e.g., *Japan*, *Microsoft*) are more or less unique, names of cities (e.g., *Springfield*) and persons (e.g., *John Smith*) are generally more ambiguous. Thus, we can expect that the distribution of entity types has a direct impact on the difficulty of the entity linking task.
We analyzed the types of entities in DBpedia with respect to our benchmark datasets. For that analysis, we used RapidMiner[12] with the Linked Open Data extension (Ristoski et al., 2015). Figure 2 shows the overall distribution, as well as a breakdown by the most frequent top classes. Although types in DBpedia are known to be notoriously incomplete (Paulheim and Bizer, 2014), and NIL entities are not considered, these figures still reveal some interesting characteristics:

- AIDA-YAGO2 has a tendency towards sports related topics, as shown in the large fraction of sports teams and athletes.
- NEEL2014 and WES2015 treat time periods (i.e., years) as entities, while the others do not.
- OKE2015 and WES2015 have a tendency towards science-related topics, as shown in the large fraction of *Scientist* and *Educational Institution* entities (most of the latter are universities).
- Wikinews/MEANTIME, without surprise, has a strong focus on politics and economics, with a large portion of the entities being of classes *Office Holder* (i.e., politicians) and *Company*.
- The WES2015 corpus has a remarkably larger set of *other* and *untyped* entities. While many corpora focus on persons, places, etc., WES2015 also expects annotations for general concepts, e.g., *Agriculture* or *Rain*.

The latter is an important finding, which shows that it is hard to build NER tools that perform well in all scenarios. While for the other benchmarks, annotations of general concepts would be punished as false positives, WES 2015 would expect them and punish their absence as false negatives.

[12] http://www.rapidminer.com/



(a) Overall distribution

(b) Distribution of person subtypes

(c) Distribution of organization subtypes

(d) Distribution of location subtypes

(e) Distribution of work subtypes

Figure 2: Distribution of entity types overall (a), as well as a breakdown for the four most common top classes person (b), organization (c), location (d), and work (e). The overall distribution depicts the percentage of DBpedia entities (a), the breakdowns depict the percentages in the respective classes.

3.3. Mention annotation characterization

When annotating entity mentions in a corpus, several either implicit or explicit decisions are being made by the dataset creators, that can influence evaluations on, and the comparison between, those datasets:

Mention boundaries: inclusion of determiners ("the pope" vs "pope"), annotation of inner or outer entities ("New York's airport JFK" vs "JFK"), tokenization decisions ("New York's" vs "New York 's"), sentence-splitting heuristics.

Handling redundancy: annotating only the first vs. annotating all occurrences of an entity.

Inter-annotation agreement (IAA): one annotator vs multiple annotators, low agreement vs high agreement.

Mention ambiguity: when is there enough context for the entity to be considered non-ambiguous?

Offset calculation: using 0 vs. using 1 as the initial identifier.

IRI vs. URI: character support for ASCII vs. Unicode.

Nested entities: does the annotation allow for annotation of multiple entity layers e.g. is 'The President of the United States of America' one entity in its entirety, two entity mentions ('President' and 'United States of America') or three ('President', 'United States of America', 'The President of the United States of America')?

In the analyzed datasets, there is only limited variety on the entity boundaries and offsets, but each dataset was generated using different annotation guidelines, resulting in major differences between types of classes annotated, and which entities are (not) to be included. The 2014/2015 NEEL annotation guidelines, for example, are based on the CoNLL 2003 annotation guidelines (which also apply to AIDA-YAGO2) but where the CoNLL guidelines consider names of fictional characters as mentions of type 'Person', the NEEL guidelines consider this as a mention of type 'Character'.

3.4. Target knowledge base (KB)

It is customary for the entity linking systems to link to cross-domain KBs: DBpedia, Freebase, or Wikipedia. Every dataset listed in Section 2. links to one of these general domain KBs. Almost all of these datasets refer to DBpedia, while AIDA-YAGO2 contains links to Wikipedia (which can easily be mapped to DBpedia) and Freebase. To evaluate entity linking on specific domains or non-popular entities, benchmark datasets that link to domain-specific resources of long-tail entities are required.

4. Discussion and Roadmap

A number of benchmark datasets for evaluating entity linking exist, but our analyses show that these datasets suffer from two main problems:

Interoperability and evaluation: Dataset evaluation and interoperability between datasets are far from trivial in practice. Existing datasets use different annotation guidelines (mention and sentence boundaries,

inter-annotation agreement, etc.) and have made arbitrary choices regarding their implementation decisions, including encoding, format and number of annotators. These differences make the interoperability of the datasets and the unified comparison between entity linking systems over these datasets difficult, time-consuming, and error-prone.

Popularity and neutral domain: Although the datasets claim to cover a wide range of topical domains, in practice they seem to share two main drawbacks: skewness towards popularity and frequency, and coverage of well-known entities from a neutral domain (Section 3.4.).

While it is not possible to find a 'one-size-fits-all' approach to creating benchmark datasets, we do believe it is possible to define directions for general improvement of benchmark datasets for entity linking.

Documentation: Seemingly trivial choices such as initial offset count, or inclusion of whitespace can make a tremendous difference for the users of a dataset. This is applicable to any decision made in the process of creation of a dataset.

Standard formats: As annotation formats are becoming more standardized, dataset creators have more incentive to choose an accepted data format, or provide script that converts the original data to one or more standardized formats.

Diversity: The majority of the datasets we analysed link to generic KBs and focus on prominent entities. To gain insights into the usefulness of entity linking approaches and understand their behavior better, we need to evaluate these on datasets characterized with high confusability, low dominance and low prominence. For this purpose, we need datasets that focus on long tail entities and different domains.

5. Conclusions and Future Work

Many research papers treat benchmarks as black boxes, making it difficult to interpret efficacy improvements in terms of the individual contributions of algorithms and data. For system evaluations to provide generalisable insights, we must understand better the details of the entity linking task that a given dataset sets forth. In this paper we have analyzed a number of entity linking benchmark datasets with respect to an array of characteristics that can help us interpret the results of proposed entity linking systems.
We have proposed directions for a generic roadmap to improve existing and future datasets. We invite the whole entity linking community to extend the discussions and join in the effort being conducted as open data / open source at: `https://github.com/dbpedia-spotlight/evaluation-datasets`.

6. Acknowledgments

This work was partially supported by European Union's 7th Framework Programme via the NewsReader Project (ICT-316404), the innovation activity 3cixty (14523) of EIT Digital (`https://www.eitdigital.eu`), by the European Union's H2020 Framework Programme via the FREME Project (644771) and and the CLARIAH-CORE project financed by NWO (`http://www.clariah.nl`).

7. Bibliographical References

Adida, B., Herman, I., Sporny, M., and Birbeck, M. (2012). RDFa 1.1 Primer. Technical report, World Wide Web Consortium, http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/, June.

AndreaGiovanniNuzzolese, AnnaLisaGentile, ValentinaPresutti, AldoGangemi, DarÃoGarigliotti, and RobertoNavigli. (2015). Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges*, page 3.

Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic Milan, and Dadzie, A.-S. (2014). Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *4^{th} International Workshop on Making Sense of Microposts*, #Microposts.

Carmel, D., Chang, M.-W., Gabrilovich, E., Hsu, B.-J., and Wang, K. (2014). ERD 2014: Entity recognition and disambiguation challenge. In *37^{th} Annual ACM SIGIR CONFERENCE*.

Doddington, G., Mitchell, A., Przybocki, M. A., Ramshaw, L., Strassel, S., and Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *4^{th} International Conference On Language Resources and Evaluation*, LREC.

Gerber, D., Hellmann, S., Bühmann, L., Soru, T., Usbeck, R., and Ngomo, A.-C. N. (2013). Real-time rdf extraction from unstructured data streams. In *The Semantic Web–ISWC 2013*, pages 135–150. Springer.

Hoffart, J., Yosef, M. A., Bordin, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust Disambiguation of Named Entities. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Ling, X., Singh, S., and Weld, D. S. (2015). Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics*, 3:315–28.

McNamee, P. (2009). Overview of the TAC 2009 knowledge base population track.

Minard, A.-L., Speranza, M., Urizar, R., na Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the newsreader multilingual event and time corpus. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Moro, A. and Navigli, R. (2015). SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *9^{th} International Workshop on Semantic Evaluation*, SemEval.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.

Paulheim, H. and Bizer, C. (2014). Improving the Quality of Linked Data Using Statistical Distributions. *Interna-*

tional Journal on Semantic Web and Information Systems (IJSWIS), 10(2):63–86.

Ristoski, P., Bizer, C., and Paulheim, H. (2015). Mining the web of linked data with rapidminer. *Journal of Web Semantics*, 35(3):142–151.

Rizzo, G., Cano Amparo E, Pereira, B., and Varga, A. (2015). Making sense of Microposts (#Microposts2015) named entity recognition & linking challenge. In *5^{th} International Workshop on Making Sense of Microposts*, #Microposts.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (2014). N3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *9^{th} Language Resources and Evaluation Conference*, LREC.

Steinmetz, N., Knuth, M., and Sack, H. (2013). Statistical analyses of named entity disambiguation benchmarks. In *Proceedings of NLP & DBpedia 2013 workshop in conjunction with 12th International Semantic Web Conference (ISWC2013), CEUR Workshop Proceedings*.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Conference on Computational Natural Language Learning*, CoNLL.

Tristram, F., Walter, S., Cimiano, P., and Unger, C. (2015). Weasel: a machine learning based approach to entity linking combining different features. In *3rd International Workshop on NLP&DBpedia*.

Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L. (2015). GERBIL: General Entity Annotator Benchmarking Framework. In *24^{th} International Conference on World Wide Web*, WWW.

Waitelonis, J., Exeler, C., and Sack, H. (2015). Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval. In *Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC2015), CEUR Workshop Proceedings*.