Remixing Entity Linking Evaluation Datasets for Focused Benchmarking

Jörg Waitelonis [a], Henrik Jürges [b] and Harald Sack [c]

[a] *Hasso-Plattner-Institute, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany*
E-mail: joerg.waitelonis@hpi.de
[b] *University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany*
E-mail: juerges@uni-potsdam.de
[c] *FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*
E-mail: harald.sack@fiz-karlsruhe.de

Abstract. In recent years, named entity linking (NEL) tools were primarily developed in terms of a general approach, whereas today numerous tools are focusing on specific domains such as e. g. the mapping of persons and organizations only, or the annotation of locations or events in microposts. However, the available benchmark datasets necessary for the evaluation of NEL tools do not reflect this focalizing trend. We have analyzed the evaluation process applied in the NEL benchmarking framework GERBIL [35] and all its benchmark datasets. Based on these insights we have extended the GERBIL framework to enable a more fine grained evaluation and in depth analysis of the available benchmark datasets with respect to different emphases. This paper presents the implementation of an adaptive filter for arbitrary entities and customized benchmark creation as well as the automated determination of typical NEL benchmark dataset properties, such as the extent of content-related ambiguity and diversity. These properties are integrated on different levels, which also enables to tailor customized new datasets out of the existing ones by remixing documents based on desired emphases. Besides a new system library to enrich provided NIF [11] datasets with statistical information, best practices for dataset remixing are presented, and an in depth analysis of the performance of entity linking annotators on special focus datasets is presented.

Keywords: Entity Linking, GERBIL, Evaluation, Benchmark

1. Introduction

Named entity linking (NEL) is the task of interconnecting natural language text fragments with entities in formal knowledge-bases with the purpose to e. g. help subsequent processing tools to cope with ambiguities of natural language. NEL has evolved to a fundamental requirement for a range of applications, such as (web-)search engines, e. g. by mapping the content of search queries to a knowledge-graph [31] or to improve search rankings [38]. By linking textual content to formal knowledge-bases, exploratory search systems as well as content-based recommender systems greatly benefit from the underlying graph structures by leveraging semantic similarity and relatedness measures [34]. Likewise, social media and web monitoring systems benefit from NEL, for e. g. by the identification of persons or companies in social media content as subject of observation or tracking. A general survey on current NEL systems has been provided by Chen et al. [30].

While the number of application scenarios for NEL is on the increase, likewise the number of different NEL approaches is evolving ranging from simple string matching techniques to complex optimization based on machine learning [25]. Most NEL approaches make use of a general solution strategy, however there is an uprising trend for specialized solutions. In [42] the authors demonstrate an approach focused on medical literature while [8] examine heritage texts with

NEL. Other approaches are focused on specific entity types, such as e.g., [7], which is applied to the domain of art. Another interesting solution is [1], which can be utilized to build domain specific NEL tools. The approach of [40] extracts semantic information from mixed media types like scientific videos. This ongoing fragmentation of types of tasks aggravates the application of generic benchmarking frameworks for NEL optimization and comparison such as GERBIL [35,29] or NERD [27,26].

With GERBIL, a NEL tool optimized for the detection of person names only might be rather difficult to compare to other NEL tools of a more general focus or specialized for another topic. However, the benchmark datasets provided with GERBIL are annotated with all types of entities including organizations, events, etc. Therefore, by using these general typed benchmarks the overall achieved results with GERBIL might only be hard to compare since the assumed person-only NEL annotator would wrongly be punished with false negatives caused by non-person annotations contained in the benchmarks. The only valid way to achieve an objective evaluation would be to manually filter a dataset to only contain persons and upload it to GERBIL for the desired experiment. However, these experiments are not reproducible, because it is neither clear or standardized, how the applied filtering was carried out, nor is the newly created filtered dataset always publicly available for further experiments. Moreover, it is not desirable to manage a plethora of different versions of filtered datasets. As of now, GERBIL deploys 14 annotators and 17 datasets, whereas these numbers are subject to constant change. For a detailed overview on annotators and datasets provided by GERBIL we refer to the official version[1]. Besides the already described problem, there are also more challenges faced by the GERBIL framework considering the recent development of new NEL approaches. For instance, it is highly desirable to be able to quantify the 'difficulty' of NEL problems presented in the different evaluation datasets, as e.g. the average degree of ambiguity, the completeness of annotations, etc.

A first attempt to cope with this problem was made by Hoffart et al. [12] by manually compiling the Kore50[2] corpus with the goal to capture hard to disambiguate mentions of entities. Another problem arises with the quality of annotations as described by [15] and [37] including e. g. annotation redundancy, inter-annotation agreement, topicality according to the evolving knowledge bases, mention boundaries, as well as nested annotations. Especially, completeness and coverage of annotations are essential measures to assess those annotation tasks (A2KB cf. [35]) where also the entity mention detection contributes to the overall results.

Since no 'all-in-one' perfect dataset has emerged in the past, which covers all the aspects sufficiently well, it would be beneficial to measure and provide dataset characteristics on document level to subsequently allow a recompilation of documents across different datasets according to predefined criteria into a customized corpus. E. g. for the already mentioned person-only annotator these measures would help to specifically select only those documents, which exhibit a significant number of person annotations providing a predefined level of 'difficulty'. Remixing evaluation datasets on document level leads to a better and more application specific focus of NEL tool evaluation while simultaneously ensuring reproducibility.

We have already introduced an extension of the GERBIL framework enabling a more fine grained evaluation and in deep analysis of the deployed benchmark datasets according to different emphases [39]. To achieve this, an adaptive filter for arbitrary entities has been introduced together with a system to automatically measure benchmark dataset properties. The implementation including a result visualization are integrated in the publicly available GERBIL framework.

In this paper, we present the following contributions: the work presented in [39] is brought up-to-date, consolidated, and furthermore extended with

- new additional dataset measures,
- a stand-alone library to enable customized remixing of datasets,
- a vocabulary to enrich NIF-based datasets with additional statistical information,
- a subset of available datasets has been reorganized to enable benchmarking according to the different dataset properties, and
- an in depth analysis of the performance of different annotators on the reorganized datasets is presented.

The paper is structured as follows: after this introductory section, measures to characterize NEL datasets are introduced in Sect. 2. Sect. 3 explains the GERBIL integration as well as the stand-alone library in detail, while Sect. 4 elaborates on the most inter-

[1] http://aksw.org/Projects/GERBIL.html

[2] https://datahub.io/de/dataset/kore-50-nif-ner-corpus

esting properties on datsets we have determined so far and presents more insights on the annotators performances on the reorganized and focused datasets. Finally, Sect. 5 concludes the paper with a summary of the presented work and an outlook on ongoing and future research.

2. Measuring NEL Dataset Characteristics

NEL datasets have already been analyzed to great extent. We consider these analyses to identify their potential shortcomings to be able to introduce characteristics and measures to establish more differentiated analyses. Ling et al. [15] have introduced the basic characteristics of 9 NEL datasets including the number of documents, number of mentions, entity types, and number of NIL annotations. Steinmetz et al. [33] went one step further with a more detailed view on the distribution of entity types including mapping coverage, entity candidate count, maximum recall, as well as entity popularity. Erp et al. [37] investigated on the overlap among datasets and introduced as new measures confusability, prominence and dominance as indicators for ambiguity, popularity, and difficulty.

In this paper, amongst others also a subset of the proposed characteristics has been integrated into the GERBIL benchmarking system. Compared to previous work, where either a theoretical only or an experimental only treatment of the problem was presented, this paper contributes a ready to use implementation by means of extending the GERBIL source code[3] and also provides a publicly available on-line service[4]. Besides the implementation of filtering the benchmark datasets according to the desired characteristics, the system instantly updates and visualizes the per annotator results including statistical summaries. The integration into GERBIL enables a standardized, consistent, extensible as well as reproducible way to analyze and measure dataset characteristics for NEL.

Building on that we also provide a stand-alone library[5] that computes the proposed metrics directly on NIF datasets. Without limiting the generality of the forgoing, the following explanations refer to the annotation (A2KB) as well as disambiguation tasks (D2KB) of the GERBIL framework. D2KB is the task of disambiguation of a given entity mention against the knowledge base. With A2KB, first entity mentions have to be localized in the given input text before the subsequent disambiguation task is performed. Hence, for most implementations D2KB can be seen as a sub task of A2KB.

Measure	Level
Not annotated	ds
Density	ds, doc
Prominence	ds, doc, an
Maximum recall	ds
Likelihood of confusion	ds, doc, an
Dominance	ds
Types	ds, doc, an

Table 1

Overview of the introduced measures and the according levels of reference, where (**ds** stands for dataset level, **doc** for document level **an** for annotation level).

To enable a more differentiated NEL evaluation, the following characteristics are introduced with the purpose to perform analysis on dataset, document, as well as entity mention level.

To define the measures the following notation is used. A dataset D is a set of documents $t \in D$. A document consists out of annotations and text $t = (T, A)$ where T is the textual representation for the document and A is the set of annotations defined on the text. We define some basic functions on the documents.

$$len : t \rightarrow \mathbb{N} \tag{1}$$

$$a : t \rightarrow \mathbb{N} \equiv a((T, A)) = |A| \tag{2}$$

The function $len(t)$ returns the number of words (whitespace separated) of a document text. The second function $a((T, A))$ returns the number of annotations within a document $|A|$. Furthermore, let E_D denote all entities within a dataset and S_D denote all used surface forms within a dataset. At last $|D|$ denotes the number of documents within a dataset.

The defined measures might refer to different levels: dataset level, document level, and annotation (or entity) level. Table 1 contains an overview on which measure is considered at a specific level. Subsequently, the introduced measures will explained in more detail.

2.1 Number of Annotations

In general, the number of annotations $|A|$ within a document is a measure to estimate the size of the disambiguation context. The average number of annotations $na(D) \rightarrow \mathbb{R}$ per document for a document corpus D equals to

$$na(D) = \frac{\Sigma_{(T,A) \in D}|A|}{|A|} \tag{3}$$

[3] https://github.com/santifa/gerbil/

[4] http://gerbil.s16a.org/

[5] https://github.com/santifa/hfts

2.2 Not Annotated Documents

Some of the available benchmark datasets even contain documents without any annotations at all. Documents without annotations lead to an increase of false positives in the evaluations and thereby cause a loss of precision. The number of not annotated documents is calculated for a document corpus D with $nad(D) \in [0,1]$:

$$nad(D) = \frac{\sum_{t \in D}(a(t) = 0)}{|D|} \tag{4}$$

Empty documents are a problem for the annotation task (A2KB), but not for the disambiguation only task (D2KB), where empty document annotations are simply omitted in the processing.

2.3 Missing Annotations (Density)

Similar to not annotated documents, missing annotations in an otherwise annotated document lead to a problem with the A2KB task. Annotators potentially identify these missing annotations, which are not confirmed in the available ground truth and thus are counted as false positives. It is not possible to determine the specific number of missing annotations without conducting an objective manual assessment of the entire ground truth data, which requires major effort. However, we propose to estimate this number by measuring an annotation density value as the relation between the number of annotations in the ground truth $a(t)$ and the overall document length $len(t)$, determined as the number of words, with $ma(D) \in [0,1]$:

$$ma(D) = \frac{\Sigma_{t \in D} a(t)}{\Sigma_{t \in D} len(t)} \tag{5}$$

If an annotation is spanning more than one word, it is only counted as one annotation.

2.4 Prominence (Popularity)

The assumption of [37] is, that an evaluation against a corpus with a tendency to focus strongly on prominent or popular entities may cause problems. Hence, NEL systems preferring popular entities potentially exhibit an increase in performance. To verify this, we have implemented two different measures on the entity level. Similarly to [37], the prominence is estimated as PageRank [21] of entities, based on their underlying link graph in the knowledge base. Additionally, we also take into account Hub and Authorities (HITS) values as a complementary popularity related score. PageRank as well as HITS values were obtained from [24].

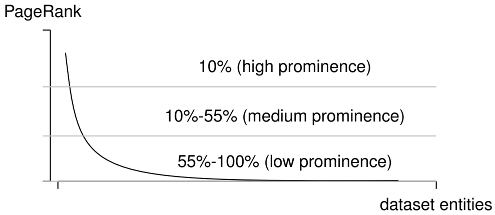

Fig. 1. Example partitioning for the PageRank.

To classify annotations, documents, and datasets according to different levels of prominence of entities, the set of entities was partitioned as follows. PageRank (respectively HITS) underlies a power-law distribution (cf. Sect. 4.2.1), meaning that only a few entities exhibit a high PageRank and the majority of entities a lower PageRank (long-tail), cf. Fig 1. Highly prominent entities are then defined as the upper 10% of the top PageRank values. The subsequent 45% (i.e. 10% – 55%) define medium prominence and the lower 45% (i.e. 55% – 100%) low prominence.

It is important to mention that for a dataset with a stronger bias towards head entities, the entities of the middle or lower segment would then be in the higher segment for a dataset with a more even distribution. Thus, when working with multiple datasets, a global partitioning including all values of all entities is preferred.

The set of entities for every category is determined for a dataset D and a scoring algorithm. We use PageRank P for demonstration and the category interval is denoted by $a, b \in [0,1]$:

$$p(D,P) = \{e \in E_D | a \leq P(e) \leq b\} \tag{6}$$

The resulting set contains all entities of a dataset that satisfies the given interval limits. A disadvantage of this approach is that entities, which do not have a score assigned, are not part of one of the resulting sets. Similarly the prominence can be determined using the HITS values or any other ranking score.

2.5 Likelihood of Confusion (Level of Ambiguity)

Since a surface form might denote multiple meanings as well as entities might be represented by different textual representatives the likelihood of confusion is a measure for the level of ambiguity for one surface form or entity. It was first proposed in [37] for surface forms. The authors pointed out that the true likelihood of confusion is always unknown due to a missing exhaustive collection of all named entities.

An example is given in the following two figures. In Fig. 2 a document with text fragment ... *Bruce* ...

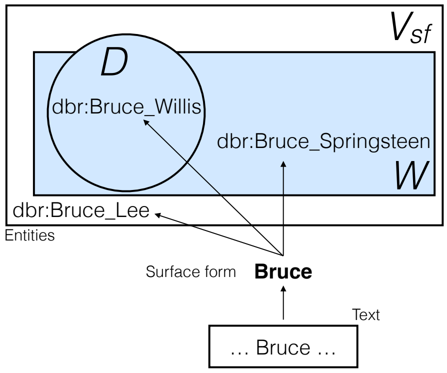

Fig. 2. The likelihood of confusion for a surface form is determined by the total number of possible entities known to some annotating system and a dataset $D \cup W_{sf}$.

that contains an entity mention is shown (lower box). The surface form 'Bruce' of the entity mention can be linked to different possible entities, i.e. they are homonyms, thus exhibiting the same writing but different meanings. The overall set of all possible entities for a surface form is V_{sf} which is also referred to as vocabulary of surface forms. The dictionary known to the annotator W_{sf} is a subset of V_{sf}. The surface forms of a dataset S_D can also be interpreted as a subset of V_{sf}. The likelihood of confusion for the surface form 'Bruce' is then determined by the cardinality of the union of the known entities $D \cup W_{sf}$, where we approximate W_{sf}. The larger the cardinality, the higher is the likelihood of confusion.

In Fig. 3 a document with text fragment *... Bruce ...* that contains an entity mention linking to the entity `dbr:Bruce_Willis` is shown. This entity could also be mapped to multiple other surface forms (synonyms). The overall set of all possible surface forms for an entity is V_e (outer lower box), which is also referred to as vocabulary of entities. The annotator knows only a subset W_e (inner lower box) of V_e, and the dataset under consideration only contains E_D, which is also a subset of V_e. *Bruce* as well as *Bruce Willis* both are surface forms used within the dataset to represent the entity `dbr:Bruce_Willis`. However, the annotation system provides *Bruce Walter Willis* as another additional possible surface form for this entity. The likelihood of confusion for an entity is then determined by the cardinality of the union of the known surface forms $D \cup W_e$.

As already shown, a surface form s or an entity e can be placed within four possible locations:

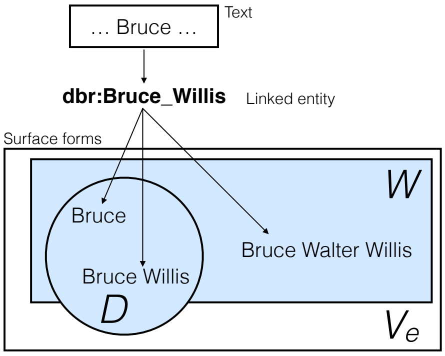

Fig. 3. The likelihood of confusion for an entity mention is the number of possible related surface forms shown in light blue.

1. Unknown to dictionary and dataset:
   $e \notin E_D \cup W_e$ or $s \notin S_D \cup W_{sf}$
2. Only known to the dataset:
   $e \in E_D \setminus W_e$ or $s \in S_D \setminus W_{sf}$
3. Only known to the dictionary:
   $e \in W_e \setminus E_D$ or $s \in W_{sf} \setminus S_D$
4. Known to dictionary and dataset:
   $e \in E_D \cup W_e$ or $s \in S_D \cup W_{sf}$

The annotator system dictionary W used for the experiments has been compiled from DBpedia entities' labels, redirect labels, disambiguation labels, and `foaf:names`, if available. For a dictionary W, the average likelihood of confusion is determined for the surface forms of a dataset S_D with $c_{sf}\colon (W, D) \rightarrow \mathbb{R}^+$. Likewise, for entities of a dataset E_D with $c_e\colon (W, D) \rightarrow \mathbb{R}^+$ is used.

$$c_{sf}(W, D) = \frac{\Sigma_{s \in S_D} e(W_{sf} \cup S_D, s)}{|S_D|} \tag{7}$$

$$c_e(W, D) = \frac{\Sigma_{e \in E_D} sf(W_e \cup E_D, e)}{|E_D|} \tag{8}$$

The function $e(W_{sf}, s)$ returns the number of entities for a surface form and $sf(W_e, e)$ returns the number of surface forms for an entity. Booth functions take also the entities or surface forms provided by the dataset into account.

Again, an annotation within a dataset contains a surface form and an entity. For each perspective (surface form or entity perspective) the likelihood of confusion is determined by counting the elements belonging to this particular perspective. For the entity perspective $c_e(W, D)$ the corresponding surface forms are used (synonyms). For the surface form perspective $c_{sf}(W, D)$ the corresponding entities are used (homonyms). The measures should roughly indicate the difficulty distribution of a dataset.

2.6 Dominance (Level of diversity)

Erp et al. introduced the dominance as a measure of how commonly a specific surface form is really meant for an entity with respect to other possible surface forms [37]. A low dominance in a dataset leads to a low variance for an automated disambiguation system and to possible over-fitting. Similar to the likelihood of confusion, the true dominance remains unknown and an approximation of the dominance is computed based on the same dictionary. In addition to the work presented in [37] we estimate dominance for both sides the entity as well as the surface form side. For an entire dataset and a dictionary, the average dominance is determined in both directions.

As e.g., for the entity `dbr:Angelina_Jolie`, let there exist 4 different surface forms in the dataset, while the dictionary provides overall 10 surface forms, which results in a 40% dominance of the entity `dbr:Angelina_Jolie` in the considered dataset. The dominance of an entity determines how many different surface forms of this entity are used in the dataset (synonyms).

As example for the other side, for the given surface form '*Anna*' the dictionary provides 10 different entities, while the dataset only uses 2 entities for different mentions with surface form '*Anna*', which results in a 20% dominance of '*Anna*' for the dataset under consideration. The dominance of a surface form determines how many different entities are used with this surface form in the dataset (homonyms). It indicates the variance or flexibility of the used vocabulary and expresses the dependency on context. Dominance indicates the expressiveness of the used vocabulary. An extensive vocabulary exhibits more diversity and is more appropriate to avoid over-fitting.

The dominance of a dataset is closely related to the likelihood of confusion since it describes the coverage among the dataset and dictionary.

The average dominance for a dataset D is determined for all entities E_D with $dom_e\colon (W, D) \to \mathbb{R}^+$ and for surface forms S_D with $dom_{sf}\colon (W, D) \to \mathbb{R}^+$.

$$dom_{sf}(W, D) = \frac{\sum_{s \in S_D} \frac{e_D(s)}{e_W(s)}}{|S_D|} \tag{9}$$

$$dom_e(W, D) = \frac{\sum_{e \in E_D} \frac{sf_D(e)}{sf_W(e)}}{|E_D|} \tag{10}$$

The function $e(s)$ returns the number of entities for a surface form and $sf(e)$ returns the number of surface forms for an entity. The index shows whether the function uses the dictionary W or the provided dataset D. Since the actual dominance is unknown and the completeness of the applied dictionaries cannot be guaranteed, computed values above the nominal threshold of 1.0 are possible. These results refer to an incomplete dictionary, i.e. there are more patterns used in the dataset than the applied dictionary does contains. The subsequently described maximum recall takes care of this aspect.

2.7 Maximum Recall

Most of the NEL approaches apply dictionaries to look up possible entity candidates matching a given surface form. If the dictionary doesn't contain an appropriate mapping for the surface form the annotator is unable to identify a possible entity candidate at all.

As Fig. 3 shows and as already mentioned before some parts of the dataset might not be contained within the dictionary. Surface forms not in the intersection are unlikely to be found by entity linking since the annotators are using dictionaries to look up potential relations. Therefore, an incomplete dictionary limits the performance of an NEL system since an unknown surface form will lead to a loss in precision. So the maximum recall can be seen as an artificial limit of a dataset.

To estimate the coverage of a mapping dictionary, the maximum recall measurement was introduced by [33]. For a dictionary W and the surface forms of a dataset S_D the maximum recall is defined as the fraction of entity mentions in the dataset and the dictionary with $max\_recall : (W, D) \to [0, 1]$:

$$max\_recall(W, D) = \frac{|\{s \in S_D | s \in W_{sf}\}|}{|S_D|}. \tag{11}$$

2.8 Types

Since some NEL approaches might be focussed on a specific domain or handle some entity categories in a different way, a filter has been implemented to distinguish dataset entities by their type. Besides the focus of NEL approaches Erp et al. also stated that types of entities may be differently difficult to disambiguate such as person names (esp. first names) might be more ambiguous and country names more or less unique [37]. For the entities of a dataset E_D, the set of entities of a specific type T is determined by $t\colon (D, T) \to (0, 1)$:

$$t(D, T) = \{e \in E_D | e \in T\}. \tag{12}$$

2.9 Micro and Macro Measurement

In accordance to Cornolti et al. [4], we distinguish between micro and macro measurements for the following measures: density, likelihood of confusion, and maximum recall. Macro measurement aggregates the average results of each single document. Regarding document length, all documents have the same influence on the aggregated result. In contrast, the micro measurement takes the results of each document into account as if they would belong to one single document, which consequently increases the influence of larger documents.

Following these theoretical considerations, the extensions of the GERBIL framework and how the determined characteristics are exploited will be described in the subsequent sections.

3. Implementation

This section describes the implementation of the GERBIL extension and the standalone library. Furthermore, the vocabulary to integrate the calculated statistics in the NIF annotation model are explained in detail.

3.1. Extending GERBIL

Two new components have been implemented to extend the GERBIL framework: one component to filter and isolate subsets of the available datasets, and a second component to calculate aggregated statistics about the data (sub-)sets according to the newly introduced measures. It is important to mention that these filters and calculations can also be applied to newly uploaded datasets. Thus, the system can also be used to gain insights about any arbitrary 'non-official' datasets not yet part of the GERBIL framework. The implemented filter-cascade is of a generic type and can be adjusted via customized SPARQL queries. E. g. to filter a dataset to only contain entities of type `foaf:Person` the following filter configuration has to be applied:

```
name=Filter Persons
service=http://dbpedia.org/sparql
query=select distinct ?v where {
    values ?v {##} .
    ?v rdf:type foaf:Person .
}
chunk=50
```

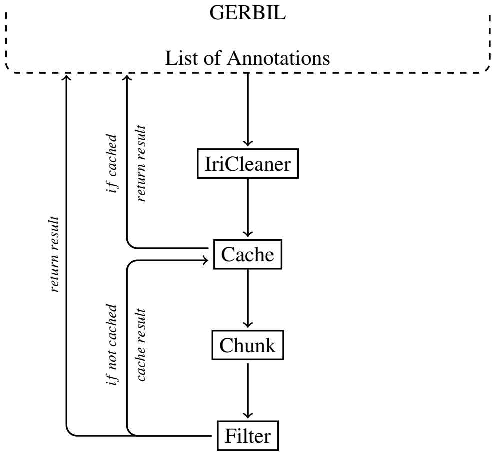

Fig. 4. Overview of the filter-cascade

The `name` designates the filter in the GUI, `service` denotes an arbitrary SPARQL-endpoint, but also a local file encoded in RDF/Turtle can be specified to serve as the base RDF query dataset. The `query` is a SPARQL query that returns a list of entities to be kept in the filtered dataset. The `##` placeholder will be replaced with the specific entities of the dataset. To avoid the size limits for SPARQL queries, the `chunk` parameter can be specified to split the query automatically in several parts for the execution. Any number of filters can be specified to be included in the analysis. With the flexibility of configuring SPARQL-queries, filters of any complexity or depth can be specified.

To partition the datasets according to entity prominence (popularity) we have additionally implemented a filter to segment the datasets in three subsets containing the top 10%, 10% to 55%, and 55% to 100% of the entities. This segmentation is applied to PageRank as well as HITS values separately.

Fig. 4 shows a general overview of the filter cascade. The annotations produced by GERBIL are subsequently cleaned from invalid IRI's. If they are already cached the result is returned. Otherwise the set is chunked und passed to the defined filter.

Buttons have been added as new control elements to the A2KB, C2KB, and D2KB overview pages in GERBIL (cf. Fig. 5). The user now is able to choose between the classic view 'no-filter', the persons, places, organisations filter views, the PageRank/HITS top 10%, 10-55%, and 55-100% filter views, a comparison view, or a statistical overview. All implemented mea-

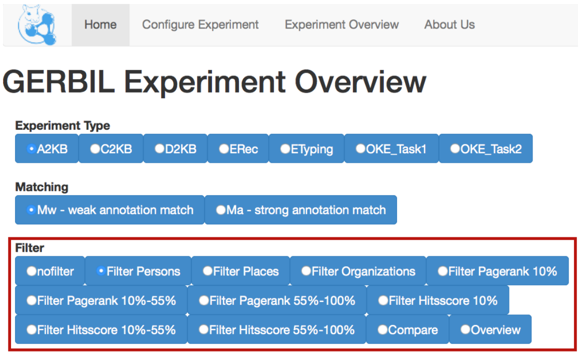

Fig. 5. New dataset filters for A2KB experiments in the GERBIL user interface.

sures are visualized in GERBIL using HighCharts[6]. The existing charts are also replaced by the new chart API, since GERBIL was limited to only one single chart type. The comparison view enables the user to view two filters at the same time as well as the average for all annotators on a specific filter. The overview shows several statistics for all datasets, such as e. g., total number of types per filter, density, likelihood of confusion in average and total. A subset of these statistics is shown and discussed in section 4. The extended source code is publicly available at Github[7]. In addition, an online version of the system is available[8].

Before discussing the dataset statistics as a result of the new GERBIL extension, the following section introduces the stand-alone-library for statistics calculation as well as the new vocabulary.

3.2. Library and Vocabulary for Dataset Statistics

Following the considerations mentioned in the previous sections, the proposed measurements can also be calculated independently of GERBIL with a separate stand-alone library. The library consumes a NIF encoded input file, calculates the proposed statistics, and extends the NIF file with the newly determined information. A comprehensive documentation as well as the library source code is provided at Github[9].

To serialize the calculated statistics generated by the GERBIL extension as well as by the library, a vocabulary has been defined with three layers to be integrated into the NIF model.

Measure	Property	Level
Not annotated	`notAnnotated`	ds
Density	`microDensity`	ds
	`macroDensity`	ds
	`density`	doc
Prominence	`hits`	en
	`pagerank`	an
Maximum recall	`microMaxRecall`	ds
	`macroMaxRecall`	ds
	`maxRecall`	doc
Likelihood of confusion	`microAmbiguityEntities`	ds
	`macroAmbiguityEntities`	ds
	`ambiguityEntities`	doc
	`ambiguityEntity`	en
	`microAmbiguitySurfaceForms`	ds
	`macroAmbiguitySurfaceForms`	ds
	`ambiguitySurfaceForms`	doc
	`ambiguitySurfaceForm`	an
Dominance	`diversityEntities`	ds
	`diversitySurfaceForms`	ds

Table 2

Overview of the introduced properties and the corresponding measurements (**ds** stands for dataset level, **doc** for document level **an** for annotation level).

The first layer refers to an entity mention, respectively annotation, (e. g. NIF phrase) with its corresponding text fragment. The second layer addresses to the document (e. g. NIF context) that provides the text where the entity mentions are embedded. A third layer groups documents together to form a dataset. We introduce the `hfts:Dataset` class, which holds the documents with the `hfts:referenceDocuments` property. On dataset level 13 properties have been introduced, which hold the measurements missing-annotation, density, maximum recall, dominance and likelihood of confusion on dataset level. Some of them come with a micro as well as macro flavour while others are only computed once.

On document level 6 new properties have been introduced to cover density, likelihood of confusion, and maximum recall. The likelihood of confusion, prominence, and the types are also assigned on entity mention level.

In Tab. 2 an overview over the introduced properties and their corresponding level is presented. Fig. 6 shows an excerpt of the extended Kore50 dataset for the new dataset class. One can see the new dataset statistics introduced by the RDF properties introduced

[6] `http://www.highcharts.com/`
[7] `https://github.com/santifa/gerbil/`
[8] `http://gerbil.s16a.org/`
[9] `https://github.com/santifa/hfts`

```
<https://.../hfts/master/ont/nif-ext.ttl/kore50-nif>
  a       hfts:Dataset ;
  hfts:diversityEntities
        "0.0661871713645466"^^xsd:double ;
  hfts:diversitySurfaceForms
        "0.08300283717687966"^^xsd:double ;
  hfts:notAnnotatedProperty "0.0"^^xsd:double ;
  hfts:referenceDocuments
    <http://.../KORE50.tar.gz/AIDA.tsv/CEL06#char=0,59> .
```

Fig. 6. An example of the new statistics properties on *dataset level* extending the KORE50 dataset.

```
<http://.../KORE50.tar.gz/AIDA.tsv/MUS03#char=0,97>
    a   nif:RFC5147String , nif:String , nif:Context ;
    nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
    nif:endIndex "97"^^xsd:nonNegativeInteger ;
    nif:isString "Three of the greatest ..."^^xsd:string ;
    hfts:ambiguityEntities "17.0"^^xsd:double ;
    hfts:ambiguitySurfaceForms "250.0"^^xsd:double ;
    hfts:density "0.17647058823529413"^^xsd:double ;
    hfts:maxRecall "1.0"^^xsd:double .
```

Fig. 7. An example of the new statistics properties on *document level* extending the KORE50 dataset.

by the *hfts:* prefix. In Fig. 7 an example for the document level is presented (`nif:Context`). Besides with the existing NIF vocabulary the statistics has been serialized with the newly introduced *hfts:* properties. The entire definition and further documentation of the vocabulary is available at Github[10].

Next, the possibility of remixing customized benchmark datasets will be explained including several examples.

3.3. *Remixing Customized Datasets*

The basic idea of remixing NEL benchmark datasets is to tailor new customized datasets from the existing ones by selecting documents based on desired emphases. This enables the compilation of focused benchmark datasets for NEL. For remixing it is proposed to store all analysed datasets in a single RDF triple store. This enables to quickly access the dataset documents via the SPARQL query language. In particular, SPARQL CONSTRUCT queries can be applied to select exactly those triples from the document annotations that meet a particular criteria, as e. g., popular persons, high possible maximum recall, places difficult to disambiguate, or any other arbitrary criteria, which can be expressed via SPARQL filter rules.

For this purpose, we introduce the basic query shown in Fig. 8. A CONSTRUCT statement creates RDF triples from document annotations meeting the filter requirement maximumRecall >= 1.0. This basic query utilizes the entire RDF induced graph and it might be useful to limit the number of documents that should be returned by the query. For this task, a subquery can be applied as shown in the second example in Fig. 9.

```
# select document triples and annotation triples
CONSTRUCT {?doc ?dPredicate ?dObject .
           ?ann ?aPredicate ?aObject .}
WHERE {
  # select all document triples
  ?ds hfts:referenceDocuments ?doc.
  ?doc ?dPredicate ?dObject .

  # select all referenced annotations
  ?ann ?aPredicate ?aObject ;
       nif:referenceContext ?doc.

  # use some filter condition
  ?doc hfts:maxRecall ?recall .
  FILTER (xsd:double(?recall) >= 1.0).
}
```

Fig. 8. Basic query that selects only documents with a maximum recall >= 1.0

Another example is presented in Fig. 10. The SPARQL subselect chooses only documents that contain persons and aggregates their number. Subsequently, the CONSTRUCT statement selects documents that contain more than 4 persons with a maximum recall of at least 0.8.

To underline that any kind of filter can be applied, Fig. 11 shows a more specific example using a federated query to select only documents from the RDF graph with persons born before 1970. To achieve this, the official DBpedia SPARQL endpoint is queried for additional information that is not present within the given benchmark datasets. More SPARQL examples can be found at Github[11].

For authoring arbitrary queries two aspects should be considered. First, many values of the proposed measurements are given as absolute values and are not always equally distributed across the datasets, documents, and annotations. Hence, it is necessary to investigate on the boundary values and value distribution before specifying a specific threshold. It is subject of future work to normalize and harmonize the statistics adequately. Second, the proposed query examples are based on document level. Therfore, if an annotation meets a requirement, the entire document together with all its annotations (which might not meet the requirement) is added to the result. Of course, queries

[10] `hfts:<https://raw.githubusercontent.com/santifa/hfts/master/ont/hfts.ttl#>`

[11] `https://github.com/santifa/hfts/blob/master/Remix.md`

```
# select document triples and annotation triples
CONSTRUCT {?doc ?dPredicate ?dObject .
           ?ann ?aPrediacte ?aObject .}
WHERE {
  # get all document triples
  ?doc ?dPredicate ?dObject .

  # limit the number of selected documents
  {SELECT DISTINCT (?d AS ?doc)
    WHERE {
      ?ds hfts:referenceDocuments ?d.
      # use this instead of a global limit
      # to ensure only documents are limited
    } LIMIT 1
  }
  # select all referenced annotations
  ?ann ?aPredicate ?aObject ;
       nif:referenceContext ?doc.

  # use some filter condition
}
```

Fig. 9. This query in addition limits the number of selected documents

```
  # document selection omitted
  ?doc hfts:maxRecall ?recall .

  # use count for a later filter expression
  {SELECT DISTINCT (?d AS ?doc) (COUNT(?a) AS ?aCount)
    WHERE {
      ?ds hfts:referenceDocuments ?d .
      # select matching entities
      ?a nif:referenceContext ?d ;
        itsrdf:taClassRef dbo:Person .
    } GROUP BY ?d LIMIT 100
  }

  # select referenced annotations omitted

  # select only documents with more than three persons
  # and a maximum recall of 0.8
  FILTER(?aCount > 3) .
  FILTER(xsd:double(?recall) >= 0.8) .
```

Fig. 10. Extract documents with a maximum recall of 0.8 and at least 4 person.

can also be structured to only return the filtered annotations, but this might lead to a missing annotation scenario that again might result in a drop of recall for the A2KB task.

Finally, the thereby newly created dataset can be uploaded to the GERBIL platform for a precisely tailored evaluation experiment.

4. Statistics and Results

This section presents the results of the execution of the proposed measures on the GERBIL datasets. Furthermore, an in depth overview on how to use the new library to partition the benchmarking datasets according to different criteria and to analyze the annotators performance in much greater detail is presented.

```
# construct block omitted
{SELECT DISTINCT (?d AS ?doc)
    WHERE {
        ?ds hfts:referenceDocuments ?d .
        # select matching entities
        ?a nif:referenceContext ?d ;
            itsrdf:taIdentRef ?ref ;
           itsrdf:taClassRef dbo:Person .

      # fetch data from another endpoint
      SERVICE <http://dbpedia.org/sparql> {
        ?ref dbo:birthDate ?date .
      }
      FILTER (?date <= xsd:date('1970-01-01')).
    }
}
```

Fig. 11. A SPARQL query that selects documents containing persons born before 1970 via additional data queried from the DBpedia SPARQL endpoint

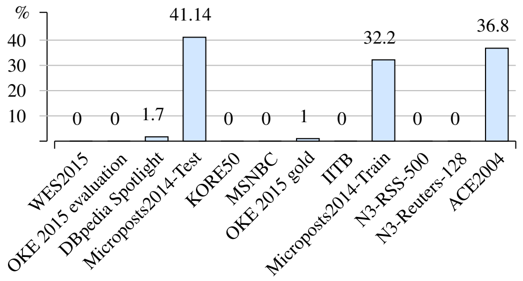

Fig. 12. Percentage of documents without annotations in the GERBIL datasets

4.1. GERBIL Datasets

The following datasets have been analyzed according to the characteristics introduced in Sect. 2: WES2015 [38], OKE2015 [20], DBpedia Spotlight [16], KORE50 [12], MSNBC [5], IITB [14], RSS500 [28], Micropost2014 [2], Reuters128 [28], and ACE2004 [18]. In this section, only the most significant results are presented. A complete listing of the achieved results is available online[12].

Fig. 12 shows the percentage of documents in the GERBIL datasets which were **not annotated**. Overall, there are 5 datasets that contain empty documents while 3 of them show a significant (i.e. >30%) number of empty documents. For A2KB tasks, these datasets will lead to an increased false positive rate and thus will lower the potentially achievable precision of an annotator. Therefore, empty documents should be excluded from evaluation datasets to enable a sound evaluation.

[12] `http://gerbil.s16a.org/`

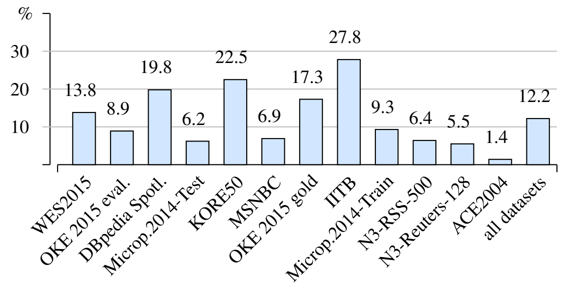

Fig. 13. Annotation density as relative number of annotations respective document length in words

Fig. 13 shows the **annotation density** of the GERBIL datasets as relative number of annotations with respect to document lengths in words. This serves as an estimation for potentially missing annotations, e. g. in the IITB dataset 27.8% of all terms are annotated. If a dataset is annotated rather sparsely (low values), it is likely that the A2KB task will result in loss of precision, because the sparser the annotations the higher is the likelihood of potentially missing annotations (as it is shown in Sect. 4.2.7). Especially for NEL tools based on machine learning it should be considered, whether a sparsely annotated dataset is appropriate for the training task. Of course, this strongly depends on the according application. Nevertheless, it is arguable, if sparseness is problematic for A2KB, because all annotators are facing the same problem and the achieved results nevertheless might still be comparable.

Table 3 shows the **distribution of entity types and entity prominence** per dataset. A green (bold) label indicates the highest value and a red (italic) the lowest value in each category. Since not all entities can be linked with a type or affiliated with the ranking, the values for each partition do not necessarily sum up to 100%. For each dataset the percentage of entities per category is denoted, as e. g., of all the entities in the KORE50 dataset 47.1% are persons and 6.9% are places. As Steinmetz et al. have demonstrated, there is a significant number of untyped entities in the DBpedia Spotlight and the KORE50 datasets. Therefore, an extra row for unspecified entities has been added to the table. The first partition (row 1–4) can be considered as an indicator of how specialized a dataset is. Thus, e. g., for the evaluation of an annotator with focus on persons, the KORE50 dataset with 45.1% of person annotations might be better suited than the IITB dataset with only 2.4% of person annotations. The second and third partition (PageRank and HITS) show the entities categorized according to their popularity. It can be observed that many datasets are slightly unbalanced towards popular entities. A well balanced dataset should exhibit a relation of 10%, 45%, 45% among the three subset categories.

Fig. 14 shows the **average likelihood of confusion** to correctly disambiguate an entity or a surface form for several datasets. The blue bar (left) indicates the average number of surface forms that can be assigned to an entity, i. e. it refers to surface forms per entity, respectively synonyms. The red/hatched bar (right) shows the average number of entities that can be assigned to a surface form, i. e. it refers to entities per surface form, respectively homonyms. The figure shows clearly that KORE50 uses surface forms with a high number of potential entity candidates, i. e. it contains a large number of homonyms. Since this dataset is focused on persons it is not surprising that surface forms representing first names, such as e. g. 'Chris' or 'Steve', can be associated with a large number of corresponding entity candidates. KORE50 was compiled with the aim to capture hard to disambiguate mentions of entities, which is confirmed by these observations. ACE2004 exposes the highest average number of surface forms for possible entities (35), i. e. it contains many synonyms.

In Section 4.2.2 a correlation analysis between likelihoods of confusion for entities and surface forms with precision and recall is presented.

Fig. 15 shows the **average dominance of entities and surface forms** in percent. The red/hatched bars show the *average dominance of entities*. The dominance of an entity expresses the relation between an entity's surface forms used in the dataset with respect to all its existing surface forms in the dictionary. Referring to Fig. 15, the KORE50 dataset uses only 9% of the surface forms that are provided in the dictionary. This indicates also how well the dataset's surface forms are covered by the dictionary's surface forms.

On the other hand, the blue bars show the *average dominance of surface forms*. The dominance of a surface form expresses the relation between of how many entities are using this surface form in the considered dataset with the overall number of entities in the dictionary using this surface form.

Referring to Fig. 15, the KORE50 dataset in which many persons are annotated uses only 7% of the possible entities for the contained surface forms. In average, entities are represented in the WES2015 dataset with 21% of their surface forms.

	WES 2015	OKE 2015 eval	DBpedia Spotl.	Microp. 2014 Test	KORE50	MSNBC	OKE 2015 gold	IITB	Microp. 2014 Train	N3-RSS-500	N3-Reuters-128	ACE2004	all datasets
Persons	18.4	30.3	3.0	16.6	**45.1**	27.2	29.3	*2.4*	16.2	15.9	6.5	6.5	18.1
Org.	3.4	11.1	3.0	9.0	16.0	9.0	18.3	*2.0*	13.8	10.5	**20.7**	20.3	11.4
Places	9.4	14.0	8.2	8.9	6.9	17.5	14.5	*3.5*	14.2	7.2	17.2	**35.0**	13.0
unspecified	68.8	44.6	85.1	65.5	**32**	46.3	37.9	*92.1*	55.8	66.4	55.6	38.2	57.4
PageRank 10%	27.9	24.4	**30.0**	21.3	28.5	28.5	24.9	14.8	26.0	*14.3*	18.8	22.2	23.5
PageRank 10%-55%	48.9	39.5	47.6	**49.8**	48.6	32.2	*0.3*	29.8	45.8	23.0	31.4	37.6	36.2
PageRank 55%-100%	22.5	16.6	19.7	**28.0**	19.4	24.8	*7.7*	15.0	25.6	11.1	19.0	15.1	18.7
HITS 10%	28.4	21.1	32.4	31.4	27.8	29.8	26.9	*12.3*	**32.9**	18.3	19.0	28.4	25.7
HITS 10%-55%	12.9	12.4	18.2	14.4	20.8	**22.8**	*0.3*	12.2	13.6	7.3	9.1	11.4	13.0
HITS 55%-100%	**58.0**	47.0	48.2	51.8	47.2	32.1	50.2	35.2	50.6	*23.2*	40.6	15.3	41.6

Table 3

Percentage of entities by entity type and entity popularity per dataset

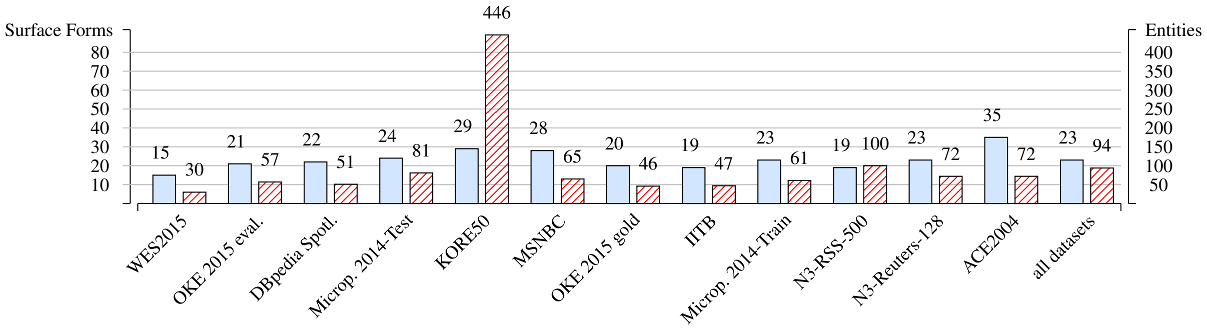

Fig. 14. Average number of surface forms per entity (blue, left) and average number of entities per surface form (red/hatched, right) indicating the likelihood of confusion for each dataset

Since the datasets with a high likelihood of confusion have a low dominance, it is arguable that these two measures express somehow the contrary. E. g. the KORE50 dataset has a high likelihood of confusion for surface forms with 446 entities for one surface form on the average. This means that for a high dominance each surface form is represented by more than 400 entities within this dataset. Such a high dominance means also that a high coverage of surface forms (dominance of entities) or entities (dominance of surface forms) is present. E. g. in the WES2015 dataset, which is focused on blog posts on rather specific topics, many rare entities (i.e. entities with a low popularity) with many different notations are used resulting in a likelihood of confusion of 15 surface forms for an entity on the average. The average dominance of entities is quite high with 21%, since the likelihood of confusion is low and topic specific blog posts often vary the surface forms for an entity to enrich the spiritedness of the text. This is commonly known from articles or essays, where the

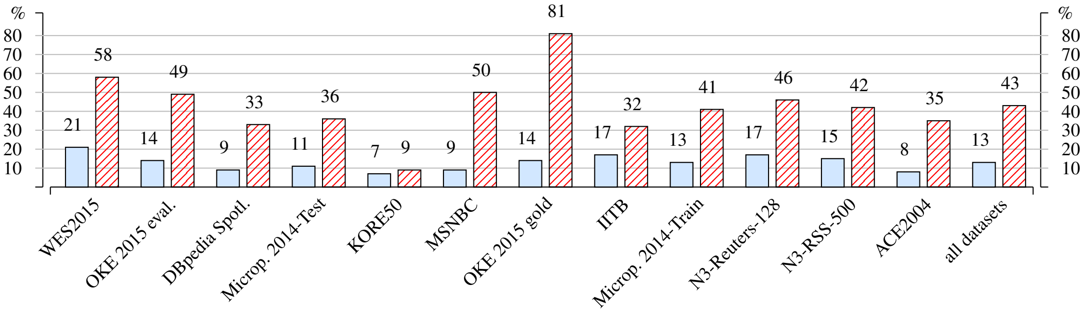

Fig. 15. Average dominance for surface forms (blue) and entities (red/hatched) per dataset

author usually tries to minimize frequent repetitions of surface form by varying the surface form for the entity under consideration to avoid monotony and to make the article more interesting to read. It might be concluded that a high dominance covers the diversity of natural language more precisely and therefore could be considered a means to prevent overfitting.

This section has introduced and discussed the results of the statistical dataset analysis. Based on these information embedded in the NIF dataset files, a cstomized reorganisation of datasets can be accomplished as explained in the following section.

4.2. *Insights from Remixing Datasets*

To gain more insights on the interplay of annotator performance and the introduced dataset characteristics, this section describes how the datasets are reorganized to determine each annotator's performance with focus on a given measure.

The approach is to first combine the datasets to one large dataset and then divide it into partitions. Each partition contains only those annotations or documents that lie in a specified interval of values of one of the proposed measures. For this purpose and to insert the statistical data into the NIF document the proposed library has been applied. Subsequently, the entire dataset was stored in an RDF triple store. With the SPARQL queries proposed in the previous sections, each partition was constructed and stored in a separate NIF document, which was submitted to the official GERBIL service to acquire the results.

For the conducted experiments the following public and GERBIL 'shipped' datasets have been used: DBpedia Spotlight, KORE50, Reuters128, RSS500, ACE2004, IITB, MSNBC. Additionally included have been the News100 [28] as well as the AQUAINT [17] dataset. Other available datasets were either not publicly available or not in the NIF format.

Since the official GERBIL service was used to conduct the experiments, the therewith provided annotators are included in the experiments. Unfortunately, not all annotators returned consistent results due to too many errors or insufficient availability. However, if sufficient results could be provided, the annotator was included in the analysis.

The following annotators provided by GERBIL have been used: AGDISTS [36], AIDA [13], Babelfy [19], Dbpedia Spotlight [16], Dexter [3], Entityclassifier.eu [6], FOX [32], Kea [41], WAT [22] and PBOH [9].

The measures used in the subsequent experiments are the measures currently supported by the library (i.e. likelihood of confusion, HITS, PageRank, density, and numbers of annotations). In general, both the A2KB as well as D2KB types of experiments, might be applied. For likelihood of confusion, HITS and PageRank only D2KB is provided because these are characteristics of the annotations. Number of annotations as estimation for the size of the disambiguation context is used with A2KB and D2KB types of tasks, density as characteristic of documents is used with A2KB only. All data as well as the achieved results can be found online[13]

4.2.1. *Value distribution and partitioning*

Fig. 16 presents the distribution of the data values over all datasets. In total, the dataset contains 16,821 annotation in 1043 documents. The figure shows a distribution chart for each measure. On the charts, the x-axis shows the number of annotations (for confusions, HITS, PageRank) or documents (for density and number of annotations). The y-axis shows the

[13] `https://github.com/santifa/hfts/blob/master/Results.md`

	Conf. Surf.		Conf. Ent.		PageRank		HITS		Num. Anno.		Density	
Part.	thr	qty	thr	qty	thr	qty	thr	qty	thr	qty	thr	qty
0	<2	8143	<2	3946	unspec.	2449	unspec.	2449	<2	20	<0.009	4
1	5	1368	3	599	<1.39E-07	3211	<5.77E-09	2456	3	595	0.015	10
2	12	1893	6	812	4.03E-07	1341	2.63E-08	19	5	63	0.023	26
3	28	2026	11	2256	1.17E-06	1504	1.20E-07	200	9	86	0.035	58
4	64	1581	19	2802	3.39E-06	2072	5.48E-07	446	16	93	0.055	194
5	147	963	34	3245	9.85E-06	2753	2.50E-06	819	29	61	0.086	333
6	338	382	62	2204	2.86E-05	1869	1.14E-05	1474	50	33	0.133	197
7	777	297	111	744	8.29E-05	1010	5.21E-05	2314	87	33	0.207	129
8	1786	128	200	203	2.40E-04	331	2.38E-04	2960	153	35	0.322	65
9	4105	40	361	10	6.98E-04	135	0.001	2744	267	24	0.500	27
10					0.002	146	0.005	940				

Table 4

Partitioning thresholds (log-based) and annotation/document quantities

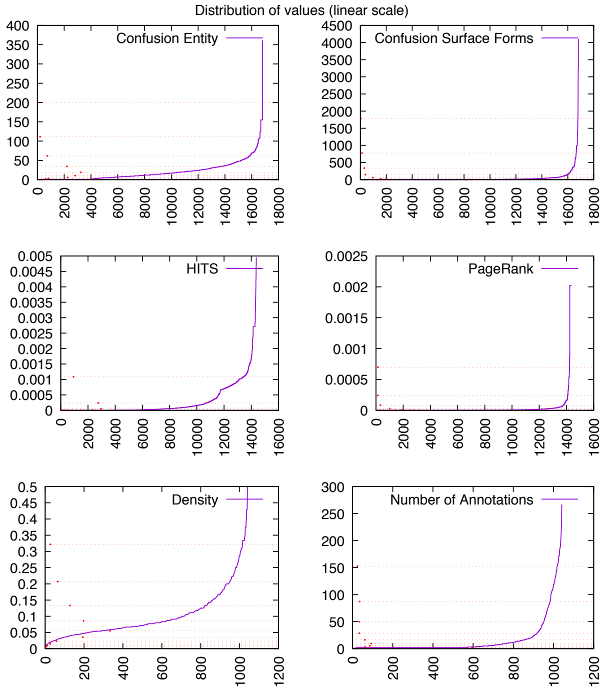

Fig. 16. Distribution of values (linear scale).

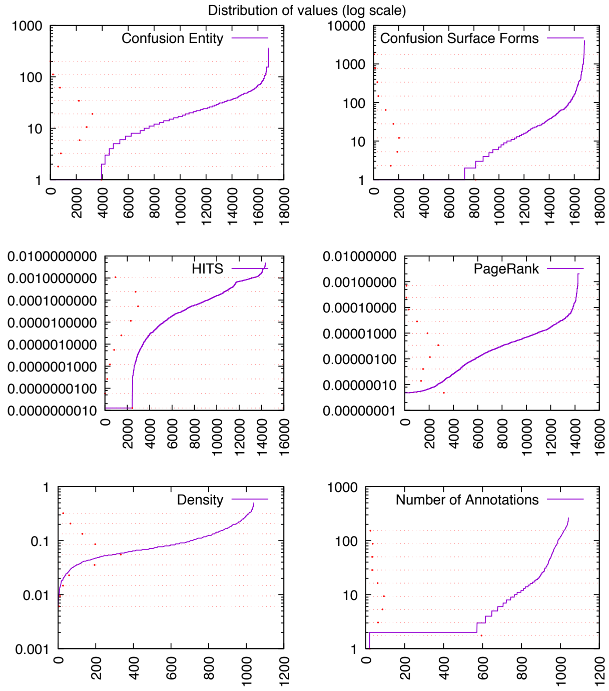

Fig. 17. Distribution of values (log scale).

absolute values of the measures. Each of the charts approximate a power-law distribution, i.e. only a few items exhibit large values and many items smaller values. For HITS and PageRank only 14,372 items are available, because for 2,449 entities no HITS or PageRank value could be determined.

We have decided to apply a decile partitioning. It seems a reasonable well choice to indicate low, medium, large as well as the boundary values. When partitioning on the item values an uneven distribution of values over the partitions occurs because of the power-law, i.e. the first partition would contain a very large disproportionate number of items and the last partition only a very few number of items. To achieve a more evenly distribution a logarithmic scaling on the values is applied as shown in Fig. 17. The red horizontal lines indicate the partition boundaries. Table 4 shows for each measure the threshold values (thr) for the partition boundaries as well as the number of items per partition (qty). For HITS and PageRank

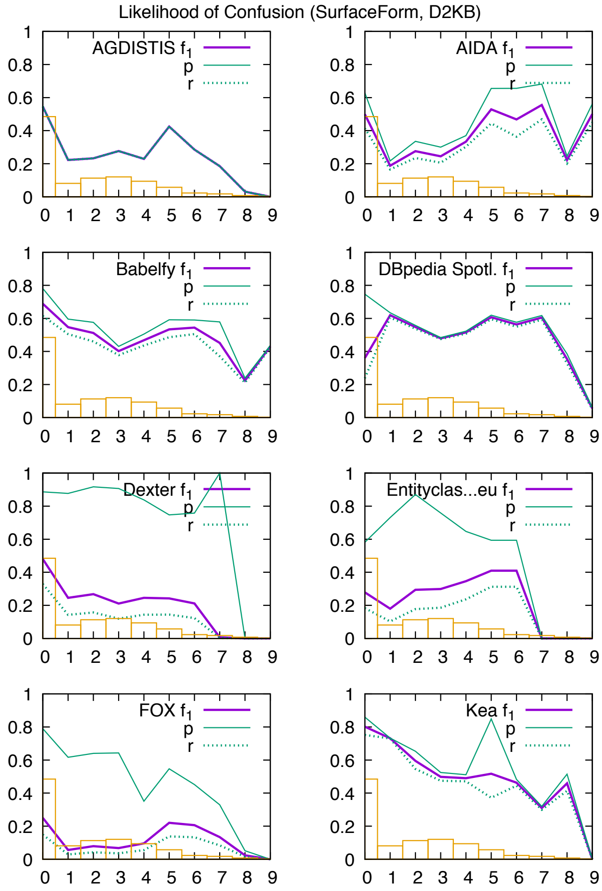

Fig. 18. Likelihood of confusion for surface forms (D2KB)

an additional partition was introduced to also include the items without a value (unspec.). Each threshold is meant as the upper boundary of the partition, thus the lower boundary is the threshold of the previous partition. The color coding will be explained subsequently.

4.2.2. *Likelihood of confusion of surface forms*

Fig. 18 shows the experimental results of each annotator for the likelihood of confusion of surface forms. Each graph shows the partitions (x-axis), as well as the determined F_1-measure (f_1), precision (p), and recall (r) for each partition. In the background the relative sizes of the partitions are indicated with boxes (see Tab. 4 for specific values).

The likelihood of confusion for surface forms describes the number of entities mapping to one particular surface form. For an annotation in the dataset, a confusion of 30 signifies that 30 possible entities for that surface form exist (homonymy).

The leftmost partition (0) contains lower values, thus annotations contain surface forms with fewer numbers of entities mapping to them and therefore a lower likelihood of confusion. Typical are for example surface forms mentioning full names, as e. g., 'Britney Spears', 'Northwest Airlines', or 'JavaScript'. The rightmost partition (9) shows larger values. It is expected that the annotations in the right partitions are more difficult to disambiguate since they exhibit a larger likelihood of confusion. The first partition contains almost half of all values, indicating that for almost half of the annotations only one entity maps to the surface form. For the second to sixth partition a reasonable even distribution is given. Considering Tab. 4, only 10 items are in the rightmost partition. These are in particular: Allen, Bill, Bob, Carlos, David, Davis, Eric, Jan, John, Johnson, Jones, Karl, Kim, Lee, Martin, Mary, Miller, Paul, Robert, Ryan, Steve, Taylor, and Thomas.

This experiment was applied as disambiguation task (D2KB)[14]. However, the entityclassifier.eu system did not provide results for partitions 7,8, and 9 (set to zero).

To interprete the figures in general, the presented graphs show a trend from the upper left to the lower right, meaning that the annotators' performance decreases with growing likelihood of confusion. Many annotators, except AIDA and Babelfy, fail with surface forms having more than ca. 1,700 entities mapping to (8th partition and above). Entityclassifier.eu , Dexter, and FOX show a very strong focus on precision, at the expense of recall, as we can also see in the further experiments.

It can be concluded that the fewer entities are mapping to a particular surface form, the easier seems the disambiguation task. For surface forms with more than 1,700 potential entity candidates the reliability of the disambiguation might drop dramatically.

4.2.3. *Likelihood of confusion of entities*

Fig. 19 shows the experimental results of each annotator for the likelihood of confusion of entities. The graphs are presented in the same way as for the previous measure. The likelihood of confusion for entities describes to how many surface forms the entity of an annotation is mapping to. For an annotation, a confusion of 30 means that 29 surface forms besides the one within the annotation share the same entity.

The leftmost partition (0) contains lower values, thus annotations with entities mapping to only one sur-

[14] `http://gerbil.aksw.org/gerbil/experiment?id=201712060006`

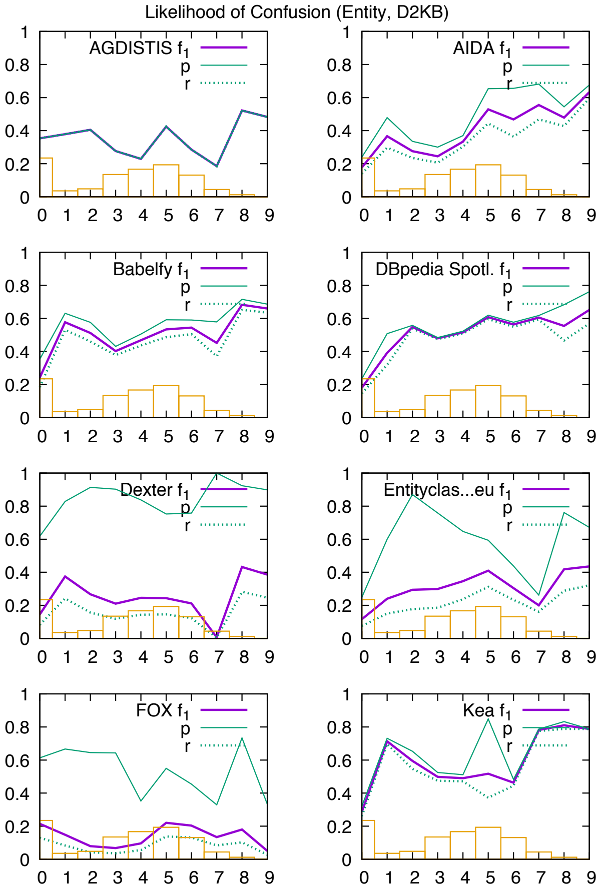

Fig. 19. Likelihood of confusion for entities (D2KB)

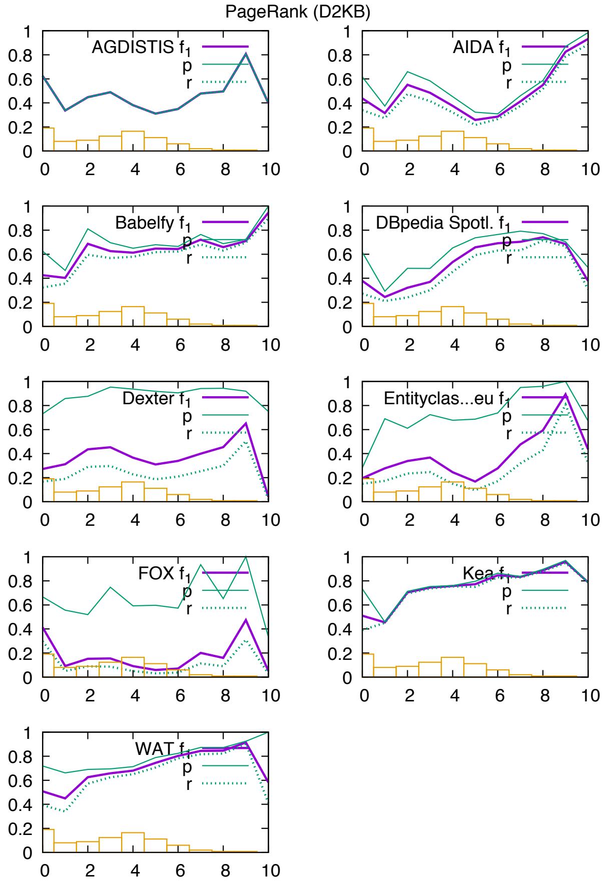

Fig. 20. Results for Pagerank (D2KB)

face form. The rightmost partition (9) contain annotations with entities mapping to more than 361 surface forms e.g. dbp:United_States. The number of items across the partitions is more evenenly distributed than for the previous measure.

This experiment was applied as disambiguation task (D2KB)[15]. Almost all participating annotators returned valid results, Entityclassifier.eu returned several faulty results.

In general, there is an upward trend, i.e., the more surface forms are available for an entity, the better it is. However, almost all annotators have in common, that the performance drops rather abruptly on the first partition (0) compared to the second partition (1). A closer look on the partition data revealed that a large share of the entities in partition 0 are resources originating from Wikipedia redirect and disambiguation pages (e. g. dbp:Diesel , dbp:Thermoelectricity). Typically, these resources only map to a single surface form, which is why they occur in partition 0. Assumably, annotators are not annotating redirect and disambiguation resources, since they prefer to use the main resource and not resources directing to it.

It can be concluded that the more surface forms an entity is mapping to, the better the annotators' performances are. Furthermore, the datasets containing a larger number of redirect and disambiguation resources can bias the annotators' performances. Future work will repeat this analysis without bias to gain insights about, how well the annotators really perform on the first partition.

4.2.4. *PageRank*

Fig. 20 shows the annotators' performances on the popularity estimation via PageRank values. Now, an additional partition is included in the graphs, which is located left (partition 0) showing the results on the

[15] http://gerbil.aksw.org/gerbil/experiment?id=201712050002

2,449 annotations, where no PageRank was given. For all other partitions, the PageRank values increase from left to right. Thus, popular entities can be found on the right hand. The distribution of values across the partitions is reasonable even.

The experiments were conducted as D2KB task[16]. With exception of Entityclassifier.eu and FOX, all annotators returned error free results. For the time of the execution of these experiments, also the WAT annotator was available.

In the graph a general uprising trend can be observed, i.e. popular entities are better disambiguated than unpopular entities, but with exception of AIDA and Babelfy, all annotators struggle with extremely popular entities (partition 10). A view in the data revealed that the 146 annotations only refer to the 4 entities `dbp:Germany`, `dbp:United_States`, `dbp:Americas` and `dbp:Animal`. Therefore, partition 10 might not be sufficiently representative. The entities with the largest PageRanks (e.g. from partition 8) mostly refer to countries and popular locations as well as to the entity `dbp:Insect`.

In conclusion, a positive correlation (>0.7) between the PageRank values and the annotator performances can be observed. It seems likely that popular entities are used much more frequently, while being described via many varying surface forms.

4.2.5. HITS

Similarly to PageRank, HITS values were not provided for all entities, thus partition 0 contains the annotations with unspecified values (see Fig. 21). For the other partitions the HITS values are increasing from left to right. According to Tab 4, partition 2 contains only very few annotations (19). The other partitions contain a more representative number of items.

Again, the experiments were conducted as D2KB tasks[17]. However, the Entityclassifier.eu annotator produced too many faulty results and had to be excluded from the evaluation.

The HITS analysis reveals that for very low values (partition 1) and higher values (partition 6 and upwards) the annotators provide better results than for the medium values (partitions 2-5). There is a weak correlation among HITS and confusion of entities (>0.4). THis could be interpreted as with increasing partition number there are less entities with lower popularity, which might cause better disambiguation results.

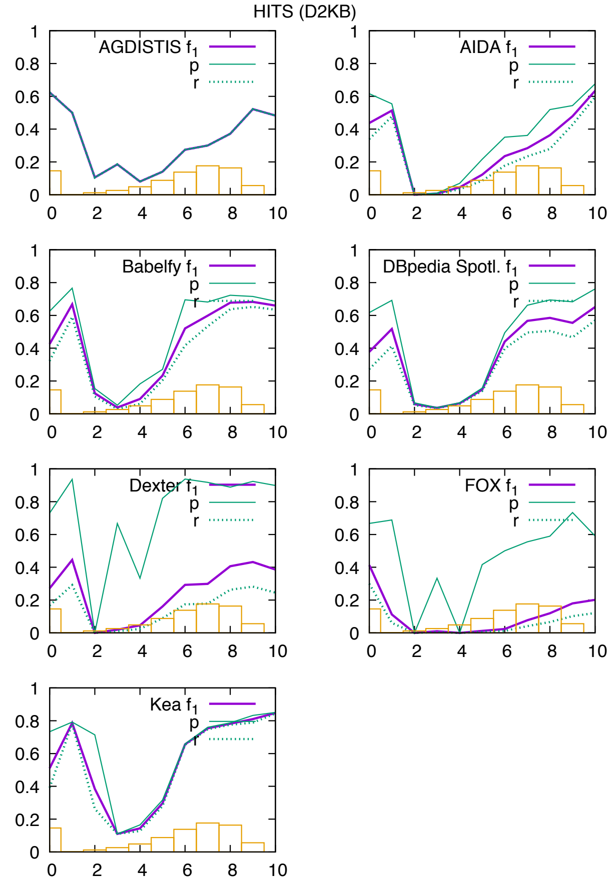

Fig. 21. Results for HITS (D2KB)

4.2.6. Number Of Annotations

Fig. 22 and 23 show the results for the number of annotations measure. This measure is not to be interpreted as a quality of the annotations but of the documents. Tab. 4 shows that more than half (595) of the 1,043 documents contain exactly 3 annotations, indicated by partition 1. Only 20 documents contain fewer annotations (partition 0). The number of annotations also corresponds to the size of the 'disambiguation context'.

For this measure both experiment types D2KB[18] (Fig. 22) and A2KB[19] (Fig. 23) were conducted. For the A2KB task, the AGDISTIS annotator was not

[16] `http://gerbil.aksw.org/gerbil/experiment?id=201712060001`

[17] `http://gerbil.aksw.org/gerbil/experiment?id=201712060011`

[18] `http://gerbil.aksw.org/gerbil/experiment?id=201711280011`

[19] `http://gerbil.aksw.org/gerbil/experiment?id=201711280030`

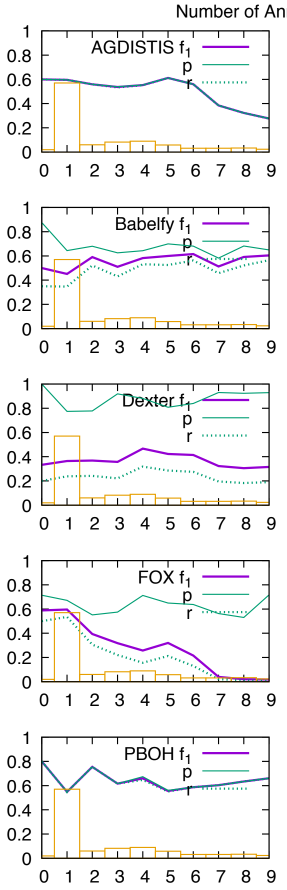

Fig. 22. Results for Number of Annotations (D2KB)

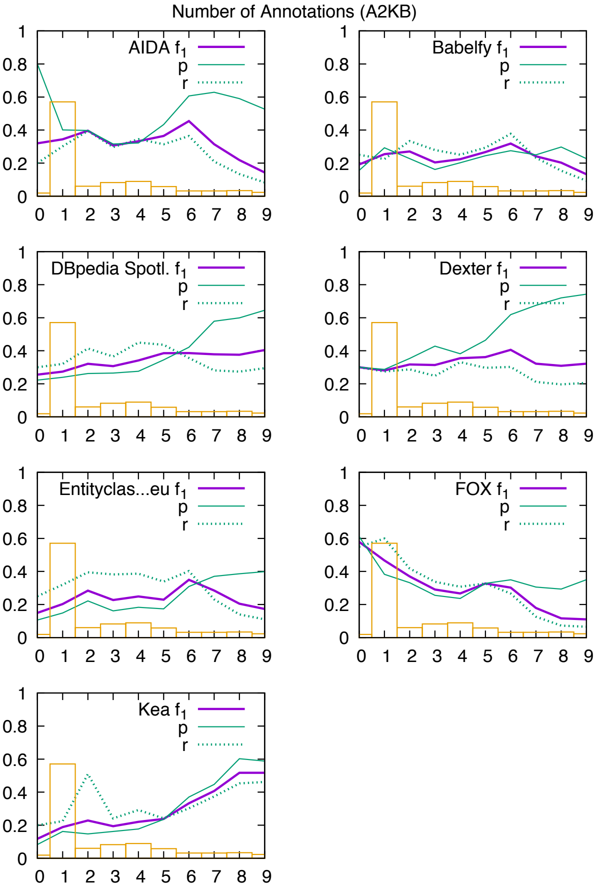

Fig. 23. Results for Number of Annotations (A2KB)

available, because it is only capable of D2KB tasks. For the period of D2KB experiments also the PBOH annotator was available. Entityclassifier.eu produced several errors, but overall, the results seem to be valid.

In Fig. 22 (D2KB) it can be observed that some annotators are not robust against growing context size, as e. g., AGDISTIS, AIDA, Entityclassifier.eu, and FOX. The other annotators exhibit a more or less constant behaviour. The annotation tasks (A2KB) presented in Fig. 22 confirm this observation. Almost every annotator increases precision with growing context sizes, but on the expense of recall. This drifting apart occurs between the 4th and 6th partition (16 to 50 annotations per document). KEA seems to strongly benefit from increasing context sizes, while FOX benefits from smaller context sizes.

4.2.7. *Density*

The results for the density measure are presented in Fig. 13. Density also is a quality of the documents and not of their annotatioms. Low density (left hand partitions) signifies that a longer document has only few annotations. High density (right hand partitions) on the other hand signifies that a document contains many annotations relative to its length.

For density the experiments were conducted as A2KB tasks[20]. All participating annotators provided valid results.

From the presented graphs it can be observed that the annotators perform on low dense documents with high recall, but comparably low precision. On the other hand, dense documents are annotated with higher precision, but lower recall. While Babelfy performs more or less evenly distributed, KEA seems to also maintain recall with denser documents. The break even point between precision and recall is located between the 4th and 6th partition (density between 0.055 and 0.133).

[20] `http://gerbil.aksw.org/gerbil/experiment?id=201712050010`

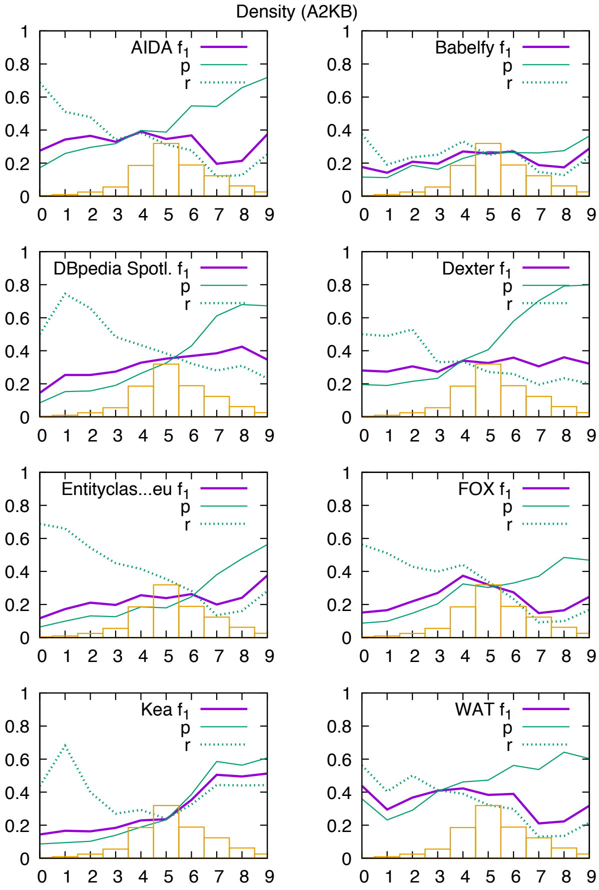

Fig. 24. Results for Density (A2KB)

4.2.8. *General Results*

Table 5 shows the achieved micro-f_1 results of the annotators for the D2KB task. The top row indicates the original GERBIL results[21] (No Filter). Top results are indicated in green (bold) and the lowest results in red (italic). Each row shows the results for the dataset filtered according to a specific criteria. The second column shows the number of remaining annotations in the dataset after filtering. The penultimate column shows the average of the annotators, the last column the Pearson correlation of the current row to the first row.

For persons[22], organizations[23] and places[24] the results achieved by the annotators are rather similar, but do not perfectly correlate to the baseline (first row). For persons and organizations PBOH seems to be the best annotator. KEA produces the best results for places and for the entities not falling into these categories (others). The others category strongly correlates with the baseline.

The next 2 rows separate annotations into a dataset containing entities with `itsrdf:taClassRef` statement (with Classes[25]) and without (without Classes[26]). The first dataset correlates very strongly to the baseline. For the annotations without class assignment the correlation is not so clear, furthermore the annotation performance was comparably low.

Another filtering was performed by filtering entities according to class membership of typical classes of the tree different domains: Music[27], Science[28], and Movie/TV[29]. In every domain a different annotator performed best. Pearson value for Music indicates a lower correlation.

The last four rows show datasets filtered according to thresholds of the proposed measures. For the first, we removed the first and last decile partition to avoid bias caused by disambiguation and redirect resources, too popular and unpopular entities, entities without information about PageRank and HITS, extremely short and large contexts, extreme homonyms and synonyms (likelihood of confusion). Furthermore, the density was restricted to a moderate level around the break even points between precision and recall to avoid major bias caused by extreme strong and low density. The filtered dataset is denoted as the 'Fair'

[21] `http://gerbil.aksw.org/gerbil/experiment?id=201711230013`

[22] `http://gerbil.aksw.org/gerbil/experiment?id=201711280013`

[23] `http://gerbil.aksw.org/gerbil/experiment?id=2017112800143`

[24] `http://gerbil.aksw.org/gerbil/experiment?id=201711280015`

[25] `http://gerbil.aksw.org/gerbil/experiment?id=201711280028`

[26] `http://gerbil.aksw.org/gerbil/experiment?id=201711280020`

[27] `http://gerbil.aksw.org/gerbil/experiment?id=201712060008` `http://gerbil.aksw.org/gerbil/experiment?id=201712110000`

[28] `http://gerbil.aksw.org/gerbil/experiment?id=201712060009` `http://gerbil.aksw.org/gerbil/experiment?id=201712110001`

[29] `http://gerbil.aksw.org/gerbil/experiment?id=201712060007`

dataset[30]. Considering Tab. 4, a grey cell background indicates that this partition was *not* included in the fair dataset. The dataset contains 765 annotations in 118 documents.

From all these restrictions, all annotations have been filtered, which fall into the intersection of the opposite filters, denoted as the 'Unfair' dataset[31] (grey cells of Tab. 4). This results in only 66 annotations in 22 documents.

Tab. 5 shows that the results for the fair dataset are overall better than for the unfair dataset. But surprisingly, 3 annotators (KEA, AGDISTIS, Dexter) perform with larger f-measure than on the fair dataset. With a larger value of 0.898 the Pearson value suggests a slightly better correlation with the baseline for the fair dataset than for the unfair dataset with 0.866.

The last two remixed datasets are a subset of the fair dataset. The first one was compiled with the intent to include only annotations, which are comparably easy to disambiguate[32]. The other one includes annotation,s which are considered more difficult to resolve[33]. Considering Tab. 4 the green, orange, and white partitions belong to the easy dataset, the red, orange and white partitions belong to the difficult datasets. We did not further restrict the number of annotations and density values compared to the fair dataset, because the result datasets would have been too small.

KEA performed well on the dataset that was considered easier, but not on the difficult dataset where PBOH is ahead of all other annotators. The average numbers of the easy and difficult datasets suggest that expectations have been fulfilled. The dataset considered more difficult to solve in fact is more difficult to solve and the easy dataset easier to solve than others. The results for the difficult dataset only slightly correlates with the overall results, but, the values for FOX are missing, so it might be not representative.

5. Conclusion

In this paper an extension of the GERBIL framework has been introduced to enable a more fine grained evaluation of NEL annotators.

According to the predefined entity types, the KORE-50 benchmark dataset contains the most persons, N3-Reuters-500 the most organizations, and ACE2004 the most places. The IITB dataset on the other hand contains almost no persons, organizations, or places. According to the PageRank algorithm the DBpedia Spotlight dataset contains the most prominent entities, while the Micropost 2014 Test dataset contains the most entities with medium and low prominence. N3-RSS contains the fewest popular and OKE 2015 gold standard the fewest medium and low prominence entities. The HITS value showed a more diverse picture with Micropost 2014 Train containing the most popular entities, MSNBC with the most medium prominence entities, and WES2015 with the most low prominence entities. On the other hand, IITB contains the fewest high prominence entities and OKE 2015 gold standard follows with the fewest medium prominence entities. N3-RSS-500 contains the fewest low prominence entities.

A stand-alone library has been introduced to enrich documents encoded in the NIF format with additional meta information. This enables researchers to remix existing NIF-based datasets according to their needs in a reproducible manner.

An exhaustive example was presented, on how to use the library to reorganize datasets according to the measures introduced earlier. Therefore, datasets were combined and partitioned to determine and visualize for each annotator correlations between a dataset property and the annotator's performance. It was ascertained that annotators fail with homonyms with a likelihood of confusion beyond ca. 1,700 entities mapping to the surface form. From the analysis on entities' likelihood of confusions, it was confirmed that redirect and disambiguation resources strongly bias the overall results. However, the overall performance increases the more surface forms an entity is mapping to. It was also shown that the PageRank of entities correlates with the annotators performance, but only up to a certain threshold. Interestingly, for the HITS measure the annotators produced poor results on low to medium, but very good results on very low and larger values. It was further shown that not all annotators are robust against a raising number of annotations in a text to disambiguate. Many annotators tend to suffer loss of recall

[30] http://gerbil.aksw.org/gerbil/experiment?id=201712100002

[31] http://gerbil.aksw.org/gerbil/experiment?id=201712100003

[32] http://gerbil.aksw.org/gerbil/experiment?id=201712120003

[33] http://gerbil.aksw.org/gerbil/experiment?id=201712120004

	|A|	Babelfy	Spotl.	Dexter	Ent.cl.	Fox	KEA	AGDI.	AIDA	PBOH	AVG	Pearson
No Filter	16821	0.572	0.485	0.349	0.285	*0.167*	**0.704**	0.407	0.374	0.625	0.441	
Person	1556	0.830	0.407	0.506	0.505	*0.268*	0.795	0.645	0.756	**0.839**	0.617	0.779
Organization	1084	0.731	0.530	0.519	0.487	*0.325*	0.732	0.675	0.756	**0.838**	0.621	0.796
Places	1477	0.702	0.643	0.643	0.695	*0.257*	**0.866**	0.693	0.809	0.856	0.685	0.763
other	12931	0.512	0.467	0.265	0.164	*0.113*	**0.651**	0.333	0.259	0.561	0.369	0.987
with Classes	11306	0.658	0.560	0.410	0.342	*0.129*	**0.807**	0.406	0.425	0.742	0.498	0.992
without Classes	5515	0.381	0.324	0.212	*0.168*	0.235	**0.467**	0.413	0.277	0.385	0.318	0.829
Music	525	0.545	0.449	0.560	0.511	*0.189*	**0.704**	0.582	0.684	0.656	0.542	0.693
Science	225	**0.797**	0.574	0.364	0.259	*0.136*	0.778	0.307	0.451	0.756	0.491	0.953
Movie/TV	305	0.631	0.367	0.406	0.379	*0.239*	0.618	0.477	0.515	**0.688**	0.480	0.871
Fair	765	0.617	0.614	0.327	0.428	*0.144*	0.646	0.361	0.500	**0.694**	0.481	0.898
Unfair	66	0.517	0.234	0.489	0.208	*0.029*	**0.760**	0.364	0.415	0.621	0.404	0.866
Easy	235	0.716	0.769	0.647	0.654	err	**0.811**	*0.566*	0.630	0.809	0.700	0.814
Difficult	98	0.601	0.421	*0.070*	0.126	err	0.194	0.071	0.552	**0.622**	0.332	0.428

Table 5

Micro-f_1 results of D2KB annotators for different remixed datasets.

with larger numbers of items to disambiguate. While FOX greatly performs on smaller contexts, KEA benefits from larger numbers of annotations in a context. Finally, the density measure shows that text with rather few annotations can promote recall and demote precision very unevenly.

Furthermore, an overall comparison of different filtered datasets was given including a focus on specific domains, as e.g., persons, organisations, places, music, science, movies/tv. Although KEA and PBOH perform well in the majority of cases, they are not necessarily the best performing annotators. Babelfy greatly performs on the science domain, thus, there are domain and dataset structure specific preferences across the annotators. Therefor, it is of major importance always to take into account the characteristics of datasets for entity linking benchmarks.

It is impossible to define how a perfect 'one for all' dataset should look like. However, we attempted to compile at least one dataset that is almost free of the apparent biasing factors ascertained from the proposed measures. To determine the 'difficulty' of a dataset, the confusion and popularity measures seem to be appropriate measures, but only in combination with moderate size of context and balanced density. Extreme outliers always should be avoided. Also redirect and disambiguation resources distort the result very much.

Further biasing factors identified in the datasets are NIL (notInWiki) annotations and the mixture of language versions of DBpedia. Both should be taken into account in further versions of this work. Unfortunately, the applied online annotators were not always available. Moreover, it is not clear what is the current development state of the annotators or how many annotators exist that are not connected to GERBIL, which might also worthwile to be included in further analysis.

Ongoing research is focused on the implementation of additional measures, such as e. g. those introduced by [10,23] and the annotators performance breakdown should also include the dominance and maximum recall measures. More datasets such as WES2015 and the Microposts series should be included in future versions.

Also, we would like to introduce difficulty levels for datasets along with new properties for annotation, which might be useful for further remixing, as e. g. a distinction of the NEL annotation for common and proper nouns, or the dependency on temporal context. The inter-annotators agreement might also be a valuable measure to be included into an evaluation.

The results of this work as well as the provided source code and the public online service enable to improve further benchmarks, to optimize annotators for a unprecedented level of detail, and the results enable to find the right tool or method for the desired annotation task.

In summary, evaluation on a more diverse as well as fine granular level will enable a better understanding of the NEL process and likewise fosters the development of improved NEL annotators.

References

[1] S. Bhatia and A. Jain. *Context Sensitive Entity Linking of Search Queries in Enterprise Knowledge Graphs*, pages 50–54. Springer, 2016.

[2] A. E. Cano, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts:(# microposts2014) Named Entity Extraction & Linking Challenge. In *CEUR Workshop Proceedings*, volume 1141, pages 54–60, 2014.

[3] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Dexter: an Open Source Framework for Entity Linking. In *Proceedings of the 6th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 17–20. ACM, 2013.

[4] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 249–260. ACM, 2013.

[5] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[6] M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-time Classification of Entities in Text with Wikipedia. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 654–658. Springer, 2013.

[7] M. Dragoni, E. Cabrio, S. Tonelli, and S. Villata. Enriching a Small Artwork Collection Through Semantic Linking. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, 2016, Proceedings*, pages 724–740. Springer, 2016.

[8] F. Frontini, C. Brando, and J.-G. Ganascia. Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. In *1st International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*, Portorož, Slovenia, June 2015.

[9] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann. Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 927–938, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[10] B. Hachey, J. Nothman, and W. Radford. Cheap and easy entity evaluation. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 464–469. ACL, 2014.

[11] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using linked data. In *International Semantic Web Conference*, pages 98–113. Springer, 2013.

[12] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. In *21st ACM International Conference on Information and Knowledge Management*, pages 545–554, New York, NY, USA, 2012. ACM.

[13] 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 Conference on Empirical Methods in Natural Language Processing*, pages 782–792. ACL, 2011.

[14] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466. ACM, 2009.

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

[16] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[17] D. Milne and I. H. Witten. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, pages 509–518. ACM, 2008.

[18] A. Mitchell, S. Strassel, S. Huang, and R. Zakhary. ACE 2004 Multilingual Training Corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1, 2005.

[19] 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.

[20] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, D. Garigliotti, and R. Navigli. Open Knowledge Extraction Challenge. In *Semantic Web Evaluation Challenge*, pages 3–15. Springer, 2015.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*, 1999.

[22] F. Piccinno and P. Ferragina. From TagME to WAT: a new Entity Annotator. In *Proceedings of the 1st International Workshop on Entity Recognition & Disambiguation*, pages 55–62. ACM, 2014.

[23] S. Pradhan, X. L. an Marta Recasens, E. H. Hovy, V. Ng, and M. Strube. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 30–35. ACL, 2014.

[24] D. Reddy, M. Knuth, and H. Sack. DBpedia GraphMeasures. Hasso Plattner Institute, Potsdam, July 2014, `http://s16a.org/node/6`.

[25] G. Rizzo, A. E. C. Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In *5th Workshop on Making Sense of Microposts at 24th Int. World Wide Web Conference*, volume 1395 of *CEUR-WS*, pages 44–53, 2015.

[26] G. Rizzo and R. Troncy. NERD: A framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76. ACL, 2012.

[27] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th Int. Conf. on Language Resources and Evaluation*. ELRA, 2014.

[28] 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 *LREC*, pages 3529–3533, 2014.

[29] M. Röder, R. Usbeck, and A.-C. Ngonga Ngomo. GERBIL's New Stunts: Semantic Annotation Benchmarking Improved.

Technical report, Leipzig University, 2016.

[30] W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, Feb 2015.

[31] A. Singhal. Introducing the knowledge graph: things, not strings. *Official Google Blog, May*, 2012.

[32] R. Speck and A.-C. N. Ngomo. Named Entity Recognition Using FOX. In *Proceedings of the 2014 International Conference on Posters &#38; Demonstrations Track*, ISWC-PD'14, pages 85–88, Aachen, Germany, 2014. CEUR-WS.

[33] N. Steinmetz, M. Knuth, and H. Sack. Statistical Analyses of Named Entity Disambiguation Benchmarks. In *Proceedings of NLP & DBpedia 2013 workshop at 12th International Semantic Web Conference*. CEUR-WS, 2013.

[34] T. Tietz, J. Waitelonis, J. Jäger, and H. Sack. Smart Media Navigator: Visualizing Recommendations based on Linked Data. In *13th Int. Semantic Web Conference, Industry Track*, pages 48–51, 2014.

[35] R. Usbeck et al. GERBIL – General Entity Annotation Benchmark Framework. In *24th World Wide Web Conf.* ACM, 2015.

[36] R. Usbeck, A.-C. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. AGDISTIS - Graph-Based Disambiguation of Named Entities using Linked Data. In *International Semantic Web Conference*, pages 457–471. Springer, 2014.

[37] M. van Erp, P. Mendes, H. Paulheim, F. Ilievski, J. Plu, G. Rizzo, and J. Waitelonis. Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for doing a better Job. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May 2016. ELRA.

[38] J. Waitelonis, C. Exeler, and H. Sack. Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval. In *NLP & DBpedia 2015 workshop at 14th Int. Semantic Web Conf.*, volume 1581, pages 33–44. CEUR-WS, 2015.

[39] J. Waitelonis, H. Jürges, and H. Sack. Don't Compare Apples to Oranges: Extending GERBIL for a Fine Grained NEL Evaluation. In *Proceedings of the 12th International Conference on Semantic Systems*, SEMANTiCS 2016, pages 65–72, New York, NY, USA, 2016. ACM.

[40] J. Waitelonis, M. Plank, and H. Sack. TIB| AV-Portal: Integrating Automatically Generated Video Annotations into the Web of Data. In *International Conference on Theory and Practice of Digital Libraries*, pages 429–433. Springer, 2016.

[41] J. Waitelonis and H. Sack. Named Entity Linking in #Tweets with KEA. In *# Microposts*, pages 61–63. CEUR-WS, 2016.

[42] J. G. Zheng, D. Howsmon, B. Zhang, J. Hahn, D. McGuinness, J. Hendler, and H. Ji. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(1):S4, May 2015.