Don't compare Apples to Oranges – Extending GERBIL for a fine grained NEL evaluation

Jörg Waitelonis
Hasso-Plattner-Institute
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
joerg.waitelonis@hpi.de

Henrik Jürges
University of Potsdam
Am Neuen Palais 10
14469 Potsdam, Germany
juerges@uni-potsdam.de

Harald Sack
Hasso-Plattner-Institute
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
harald.sack@hpi.de

ABSTRACT

In recent years, named entity linking (NEL) tools were primarily developed as general approaches, 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 used 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 [16] and its benchmark datasets. Based on these insights we extend the GERBIL framework to enable a more fine grained evaluation and in deep analysis of the used benchmark datasets according to different emphases. In this paper, we present the implementation of an adaptive filter for arbitrary entities as well as a system to automatically measure benchmark dataset properties, such as the extent of content-related ambiguity and diversity. The implementation as well as a result visualization are integrated in the publicly available GERBIL framework.

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 better deal 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 [13] or to improve search rankings [18]. When 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 or relatedness measures [15]. Social media and web monitoring systems are supported by 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 NEL systems is given by Chen et al. [12].

SEMANTiCS 2016, September 12-15, 2016, Leipzig, Germany

DOI: http://dx.doi.org/10.1145/2993318.2993334

While the number of application scenarios for NEL is on the increase, likewise the number of different approaches is evolving ranging from simple string matching techniques to complex optimization via machine learning [8]. Most NEL approaches usually follow a generic solution strategy, but there is an uprising trend for many systems focussing on the solution of rather specific tasks only, e. g. by the restriction to a specific domain of interest, document-, or entity type. This ongoing fragmentation of types of tasks aggravates the application of generic benchmarking tools for NEL optimization and comparison such as GERBIL [16, 11], which is based on the BAT-framework [1], or NERD [10, 9]. With GERBIL, a NEL tool optimized for the detection of person names only is difficult to compare to other NEL tools which have a more general focus. The datasets provided with GERBIL are annotated with all types of entities including organizations, events, etc. Hence the overall achieved results with GERBIL are not comparable since the person only NEL annotator would wrongly be punished with false negatives caused by the contained non-person annotations. 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 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 we refer to the official version[1].

Besides the already described problem, there are other 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.

A first attempt was made by Hoffart et al. [3] by manually compiling the Kore50[2] corpus aiming to capture hard to disambiguate mentions of entities. Another problem arises with the quality of annotations as described by [4] and [17] including e. g. annotation redundancy, inter-annotation agreement, topicality according to the evolving knowledge-bases, mention boundaries and nested annotations. Especially, completeness and coverage of annotations are essential measures when assessing the annotation tasks (A2KB cf. [16]) where

[1] http://aksw.org/Projects/GERBIL.html
[2] https://datahub.io/de/dataset/kore-50-nif-ner-corpus

also the entity mention detection contributes to the overall results.

Since no 'all-in-one' perfect data-set 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 re-compilation of documents across different datasets according to predefined criteria. E.g. for the already mentioned person only annotator these measures would help to specifically select only those documents, which exhibit a significant amount of person annotations providing a specific 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.

In this paper we introduce an extension of the GERBIL framework enabling a more fine grained evaluation and in deep analysis of the used benchmark datasets according to different emphases. To achieve this, we have implemented an adaptive filter for arbitrary entities as well as a system to automatically measure benchmark dataset properties. The implementation as well as a result visualization are integrated in the publicly available GERBIL framework, building a fundamental requirement to be able to remix and customize NEL evaluation data.

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 in detail, while Sect. 4 elaborates on the most interesting results we have achieved so far. Finally, Sect. 5 concludes the paper with a summary and an outlook on ongoing and future research.

2. MEASURING NEL DATASET CHARACTERISTICS

NEL datasets have been analyzed to great extent. We consider these analyses to identify their potential shortcomings to be able to introduce characteristics and measures to enable more differentiated analyses. Ling et al. [4] have introduced the basic characteristics of nine NEL datasets including the number of documents, number of mentions, entity types, number of NIL annotations. Steinmetz et al. [14] went further with a more detailed view on the distribution of entity types as well as mapping coverage, entity candidate count, maximum recall, and entity popularity. Erp et al. [17] investigated on the overlap between datasets and introduced confusability, prominence and dominance as indicators for ambiguity, popularity, and difficulty.

In this paper, besides others, the implementation of a subset of the proposed characteristics as an integration into the GERBIL benchmarking system is introduced. Compared to previous work, where a theoretical only as well as experimental only treatment of the problem is presented, this paper contributes a ready to use implementation by means of extending the GERBIL source code[3] and also providing an 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 as well as statistical summaries. The integration in GERBIL enables a standardized, consistent, extensible as well as reproducible way of measuring dataset characteristics for NEL.

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 disambiguating a given entity mention against the knowledge base. With A2KB, first the entity mentions have to be localized in the given input text, before the disambiguation task is performed. Hence, for most implementations D2KB can be seen as a sub task of A2KB.

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.

2.1 Not Annotated Documents

Some of the datasets 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. For a dataset D, documents $t \in D$ and the set of annotations $a(t)$ within t, the relative number of documents without any annotation at all $e\colon \mathcal{D} \rightarrow [0,1]$ is determined as:

$$e(D) = \frac{|\{t \in D : a(t) = \emptyset\}|}{|D|} \tag{1}$$

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.

2.2 Missing Annotations (Density)

Similar to not annotated documents, missing annotations in an otherwise annotated document lead to a problem with the A2KB task. Annotators might 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 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 and the overall document length $len(D)$, determined as the number of words, with $d\colon \mathcal{D} \rightarrow [0,1]$:

$$d(D) = \frac{\Sigma_{t \in D}|a(t)|}{\Sigma_{t \in D} len(t)} \tag{2}$$

This measure is specified as a micro measure, in example longer documents might have more influence then short documents. Moreover, if an annotation is spanning more than one word, it is only counted as one annotation.

2.3 Prominence (Popularity)

The assumption of [17] is, that evaluation against a corpus with a tendency to focus strongly on prominent or popular entities may cause problems. Hence, NEL systems preferring popular entities might exhibit an increase in performance. To verify this, we have implemented two different measures on entity level. Similarly to [17], the prominence is estimated as PageRank [5] of entities, based on their underlying link graph. Additionally, we also take into account

[3] https://github.com/santifa/gerbil/
[4] http://gerbil.s16a.org/

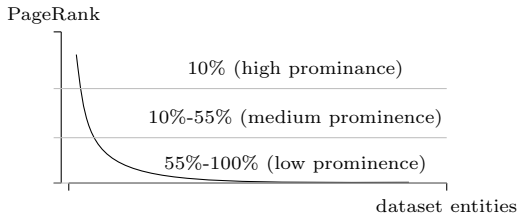

Figure 1: Example partitioning for the PageRank.

Hub and Authorities (HITS) values as an additional popularity related score. PageRank as well as HITS values were obtained from [7].

To evaluate annotators according to different levels of prominence of entities, the set of entities was partitioned as follows. A power-law distribution of the PageRank (respectively HITS) values over all entities is assumed, meaning that only a few entities exhibit a high PageRank and many 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.

For a dataset, the relative amount of entities for every category is determined with $pr\colon (\mathcal{D}, \mathcal{PR}) \rightarrow [0,1]$ using the PageRank PR and the interval $a, b \in \mathbb{R}$ where e refers to a single entity:

$$pr(D, PR) = \frac{|\{e|\forall e \in D, a \leq e \leq b \wedge e \in PR\}|}{|e \in D|} \quad (3)$$

The resulting set contains all entities of a dataset that satisfies the given interval limits. Similarly the prominence can be determined using the HITS values or any other ranking score.

2.4 Likelihood of Confusion (Level of ambiguity)

Since a surface form can have multiple meanings and entities can have multiple 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 [17] for surface forms. The authors point out that the true likelihood of confusion is always unknown due to a missing exhaustive collection of all named entities. However, we apply a dictionary containing a mapping between surface forms and entities and vice versa. This dictionary has been compiled from DBpedia entities' labels, redirect labels, disambiguation labels, and 'foaf:names' if available. For an entire dataset and a dictionary W, the average likelihood of confusion is determined for surface forms S with $c_{sf}\colon (\mathcal{W}, \mathcal{S}) \rightarrow \mathbb{R}$ and entities E with $c_e\colon (\mathcal{W}, \mathcal{S}) \rightarrow \mathbb{R}$:

$$c_{sf}(W, S) = \frac{\Sigma_{s \in S}|\{e|s \in W\}|}{|S|} \quad (4)$$

$$c_e(W, E) = \frac{\Sigma_{e \in E}|\{s|e \in W\}|}{|E|} \quad (5)$$

Since the dictionary is a multi set, the term $\{x|y \in W\}$ refers to the set containing all elements x that are referenced by a search variable y. The likelihood of confusion gives only a rough overview of how difficult it might be to correctly disambiguate the entities and surface forms used in dataset. $c_{sf}(W, S)$ can also be seen as an indicator for homonyms, and $c_e(W, E)$ as an indicator of synonyms.

2.5 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 [17]. 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 in [17] 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.

In the one direction the amount of surface forms used for one specific entity in the dataset $e(D)$ is divided by the amount of possible surface forms referencing that entity in the dictionary $e(W)$. For example, for the entity `dbp: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 `dbp:Angelina_Jolie` in the considered dataset. Again, the dominance of an entity determines how many different surface forms of this entity are used in the dataset (synonyms). It indicates the expressiveness of the used vocabulary. An extensive vocabulary exhibits more diversity and is more appropriate to avoid over-fitting.

In the other direction we divide the amount of all entities for one specific surface form used within the dataset $s(D)$ by the possible number $s(W)$ referenced in the dictionary. For example, 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. Again, 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 dependence on context.

The average dominance for an entire dataset is computed over all entities $e \in D$ with $dom_e\colon (\mathcal{W}, \mathcal{D}) \rightarrow \mathbb{R}$ and surface forms $s \in D$ with $dom_{sf}\colon (\mathcal{W}, \mathcal{D}) \rightarrow \mathbb{R}$:

$$dom_e(W, D) = \frac{\Sigma_{e \in D} \frac{e(D)}{e(W)}}{|e \in D|} \quad (6)$$

$$dom_{sf}(W, D) = \frac{\Sigma_{s \in E} \frac{s(D)}{s(W)}}{|s \in D|} \quad (7)$$

Since the real dominance is unknown and the completeness of the used dictionaries cannot be guaranteed, computed values above 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.

2.6 Types

Since different NEL approaches focus on different categories of entities, we have implemented a filter to considered the following DBpedia entity types separately: person, places,

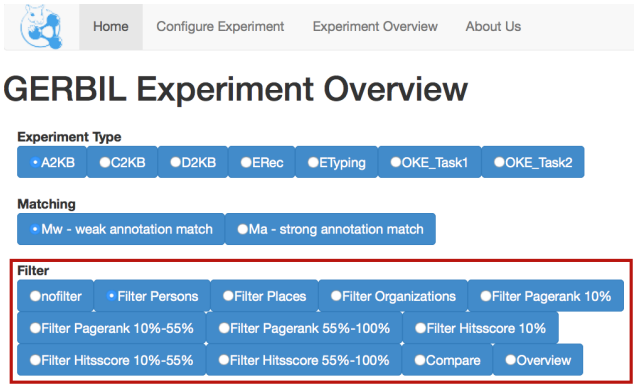

Figure 2: New dataset filter buttons for A2KB experiments.

organizations, and others. 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 might be more ambiguous and country names more or less unique [17]. For a dataset, the relative amount of entities of a specific type T is determined with $t\colon (\mathcal{D}, \mathcal{T}) \rightarrow [0, 1]$:

$$t(D, T) = \frac{|\{e | e \in D \wedge e \in T\}|}{|e \in D|} \tag{8}$$

Following these theoretical considerations, the extensions of GERBIL and how these characteristics are used will be described in the subsequent section.

3. EXTENDING GERBIL

Two new components have been implemented to extend the GERBIL framework: one component to filter and isolate subsets of the datasets, and another component to calculate aggregated statistics about the data (sub-)sets according to the 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 get insights about arbitrary 'non-official' datasets.

The filter-cascade is generic and can be arbitrarily adjusted with customized SPARQL queries. E. g. to filter a dataset to only contain entities of type `foaf:Person`, the following filter configuration can be applied:

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

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.

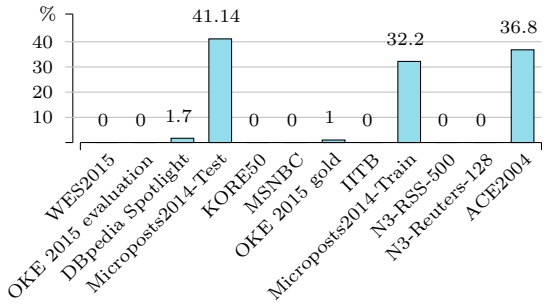

Figure 3: Percentage of documents without annotations in a dataset

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.

We have added new buttons to the A2KB, C2KB, and D2KB overview pages in GERBIL (cf. Fig. 2). The user can now 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 as well as a statistical overview.

All implemented measures are visualized in GERBIL using HighCharts[5]. The existing charts are also replaced by the new chart API, because GERBIL was limited to only one 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 amount of types per filter, density, likelyhood of confusion in average and total. The extended source code can be found on Github[6] and also an online version is available[7].

The following section will introduce a selection of the most interesting results we have determined so far.

4. RESULTS

The datasets and annotators have been analyzed according to the characteristics introduced in Sect. 2. In this section, only the most significant results are presented. A complete listing of the achieved results is available online.

Fig. 3 shows the **percentage of empty documents** in a dataset. Overall, there are six datasets that contain empty documents while four of these show a significant (>30%) amount 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 for a proper evaluation.

Fig. 4 shows the **annotation density** of the datasets as relative number of annotations with respect to documents lengths in words. This serves as an estimation for potentially missing annotations, e. g. in the IITB dataset 27.8% of

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

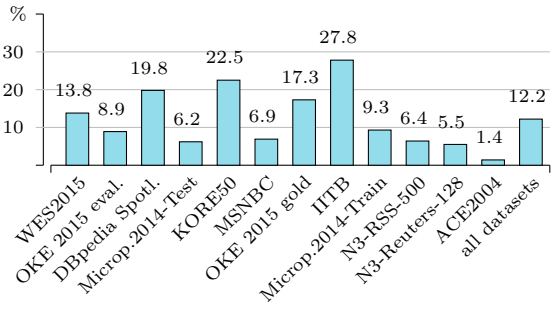

Figure 4: Annotation density as relative number of annotations respective document length in words

words are annotated. If a dataset is annotated very 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. In order to find evidence for this correlation, we have determined the Pearson correlation between density and achieved precision with a result of 0.7, which supports our original assumption. Especially for NEL tools based on machine learning it is of importance, if a sparsely annotated dataset is appropriate for the training task. Of course, this strongly depends on the application. Nevertheless, it is arguable, if sparseness is problematic for A2KB, because all annotators are facing the same problem and results might be still comparable.

Table 1 shows the **distribution of entity types and entity prominence** per dataset. An olive label indicates the highest value and a red 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 in a category is denoted, 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 are many untyped entities in the DBpedia Spotlight and KORE50 datasets. Therefore, an extra row for unspecified entities has been added. The first partition (row 1–4) can be used as an indicator of how specialized a dataset is. Thus, e. g., for the evaluation an annotator with a focus on persons, the KORE50 dataset with 45.1% of person annotations might be more efficient 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.

Table 2 shows the achieved micro-f_1 **results of the annotators** for the D2KB task, partitioned in the same way as table 1. The top row indicates the original GERBIL results (No Filter). Top results are indicated in olive and the lowest results in red. Each row indicates an entity restriction either by entity type or by entity popularity measure (PageRank and HITS) being applied before evaluation. For persons, organizations and places the results achieved by the annotators are rather similar, except for FOX, Entityclassifier.eu, and Dexter, which achieved significantly lower scores. DBpedia Spotlight performs best for places and KEA for persons as well as for organizations. Developers might use these results to optimize their systems accordingly. For example, the KEA system could be improved by investigating on why places are not sufficiently well recognized and linked.

The second and third partition shows the results achieved for entities of different popularity according to PageRank and HITS. The subsets for high, medium and low popularity show that all annotators achieve rather similar results for each subset. This observation is further supported by the average PageRank and HITS values denoted in the last column. There is no significant difference in the achieved results for popular entities vs. less popular entities. More detailed results for each possible filter and dataset as well as the results for the A2KB tasks can be obtained online.

Fig. 5 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 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 potentially possible entities, i. e. it contains many 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 huge number of corresponding entity candidates. KORE50 was made with the aim to capture hard to disambiguate mentions of entities, which is also reflected by these numbers. ACE2004 exposes the highest average number of surface forms for possible entities (35), i. e. it contains many synonyms.

To measure a correlation between likelihoods of confusion for entities and surface forms with precision and recall, the following Pearson correlation values have been determined: entity-recall = 0.156, entity-precision = -0.858, surface-recall = 0.126, and surface-precision = -0.351. Carefully speaking, these results indicate negative correlations for precision, which was expected. Thus, the more potential candidates entities exist for each surface form in a dataset (homonyms), the lower is the achieved precision. Likewise, the more different potential surface forms exist for the entities in a dataset (synonyms), the lower is precision.

With regard to recall, only a very slight positive correlation can be observed, which does not allow to draw a clear conclusion. Furthermore, the stated values do not include the KORE50 dataset, which was excluded as outlier since it exposes a very large number of homonyms within a rather small dataset only.

Fig. 6 shows the **average dominance of entities and surface forms** in percent. The blue 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. 6, the ACE2004 dataset uses only 8% of the surface forms existing in the dictionary. It indicates also how well the dataset's surface forms are covered by the dictionary's surface forms.

On the other hand, the red bars show the *average dominance of surface forms*. The dominance of a surface form expresses the relation between of how many entities are using

	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 1: Percentage of entities by entity type and entity popularity per dataset

	Babelfy	DBpedia Spotl.	Dexter	Entityclassifyer.eu	FOX	KEA	TagMe 2	WAT	AGDISTIS	average
No Filter	0.53	0.56	0.39	0.33	0.32	0.63	0.59	0.58	0.52	0.49
Persons	0.81	0.69	0.53	0.57	0.44	0.84	0.77	0.80	0.74	0.69
Org.	0.71	0.83	0.65	0.75	0.55	0.88	0.79	0.80	0.77	0.75
Places	0.77	0.82	0.57	0.55	0.54	0.78	0.81	0.80	0.75	0.71
PageRank 10%	0.68	0.76	0.50	0.48	0.39	0.79	0.74	0.75	0.63	0.64
PageRank 10%-55%	0.69	0.75	0.50	0.50	0.40	0.80	0.75	0.74	0.62	0.64
PageRank 55%-100%	0.72	0.70	0.48	0.46	0.36	0.81	0.74	0.75	0.63	0.63
HITS 10%	0.67	0.78	0.48	0.48	0.40	0.82	0.74	0.74	0.62	0.64
HITS 10%-55%	0.69	0.74	0.51	0.52	0.40	0.79	0.75	0.75	0.64	0.64
HITS 55%-100%	0.68	0.69	0.48	0.47	0.36	0.79	0.74	0.73	0.61	0.62

Table 2: Micro-f_1 results of D2KB annotators by entity type and entity popularity per dataset

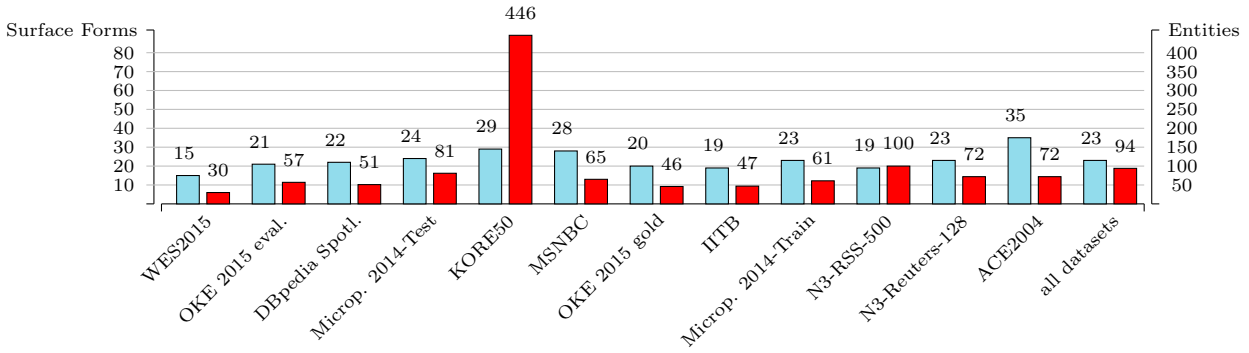

Figure 5: Average number of surface forms per entity (blue, left) and average number of entities per surface form (red, right) indicating the likelihood of confusion for each dataset

this surface form in the considered dataset with the overall number of entities in the dictionary using this surface form.

Referring to Fig. 6, the KORE50 dataset in which many persons are annotated uses only 9% of the possible entities for the contained surface forms. In average, entities are represented in the WES2015 dataset with 21% of their surface forms.

To verify a correlation between dominance for entities and surface forms with precision and recall, the following Pearson correlation values have been determined: entity-precision = 0.056, entity-recall = 0.063, surface-precision = -0.095, surface-recall = 0.674. Only the surface-recall relation shows a potential positive correlation. That means, to enable an improved recall, the surface form dominance of the datasets should be increased. Again, because of the diversity of the datasets and only scarcely available data points, these numbers are rather vague and only enable a tentative insight.

Since the datasets with a high likelihood of confusion have a low dominance, it is arguable that these two measures are somehow 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 for a high dominance each surface form is represented by more than 400 entities within the dataset. Such a high dominance means also that a high coverage of surface forms (dominance of entities) or entities (dominance of surface forms) is given. E. g. in the WES2015 dataset, which is focused on blog posts on rather specific topics, many rare entities 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 are ideal to vary the surface forms for an entity. This is commonly known from articles or essays, where the author usually tries to minimize surface form repetitions by varying the surface form for that entity to make the article more interesting to read. It might be concluded that a high dominance covers the natural language more precisely and therefore could be considered a means to prevent overfitting.

5. CONCLUSION

In this paper an extension of the GERBIL framework has been introduced to enable a more fine grained evaluation of NEL annotators. It was shown that not all of the available datasets are equally suitable for the A2KB task. According to our evaluation, the best suited datasets for A2KB are WES2015, OKE 2015 evaluation, DBpedia Spotlight, KORE50 and IITB. We have also shown that the general assumption that for annotators popular entity annotations are easier to distinguish does not hold for the considered datasets and annotators. The average scores achieved by each annotator for different levels of entity popularity are almost identical.

According to our predefined entity categories, KORE50 contains the most persons, N3-Reuters-500 the most organizations, and ACE2004 the most places. The IITB dataset on the otherhand 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. As a result, users might chose the best suited annotator for specific texts according to the properties of the considered texts.

We have documented that some of the presented measures directly correlate to precision and recall. Since there are only a few data points available and the datasets exhibit a strong variation, the correlation numbers should be considered with caution. The results of this work as well as the provided source code and the public online server 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.

Ongoing research is focussed on the implementation of a document level remixing of datasets according to different characteristics. This will enable a more focused evaluation with regard to specific applications and needs. Moreover, it would also be possible to compile a well or perfectly balanced dataset for general purpose annotators. It may also be possible to define an overall difficulty measure for given datasets

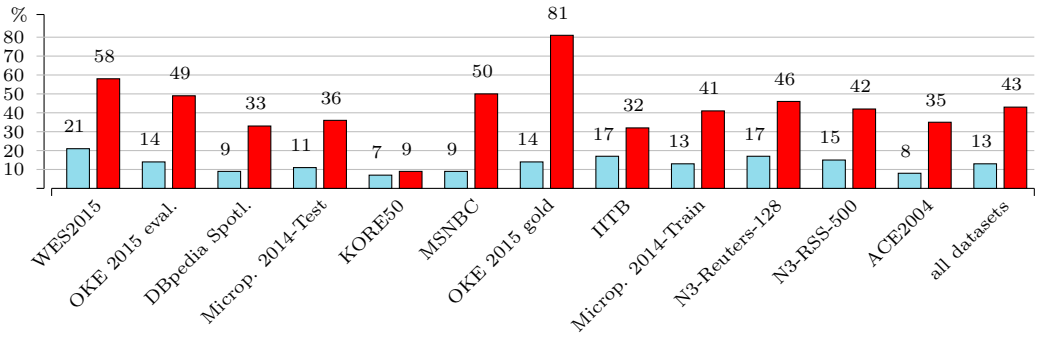

Figure 6: Average dominance for surface forms (blue) and entities (red) per dataset

or remixes. Furthermore, remixing of evaluation datasets based on a context layer is subject of future research, where only sentences with annotations fulfilling a specific measure or filter are considered to enable a dataset remixing based on themes or topics.

Overall it would be beneficial to also extend GERBIL with additional evaluation measures, such as e. g. those introduced by [2, 6] including NIL analysis or the maximum achievable recall for a given mapping dictionary [14]. In summary, evaluation on a more diverse as well as fine granular level will enable a better understanding of the NEL process and likewise foster the development of improved NEL annotators.

6. REFERENCES

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

[2] 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.

[3] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *21st ACM Int. Conf. on Information and Knowledge Management*, pages 545–554, New York, NY, USA, 2012. ACM.

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

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

[6] S. Pradhan, X. Luo, M. 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.

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

[8] 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.

[9] G. Rizzo and R. Troncy. NERD: A framework for unifying named entity recognition and disambiguation web extraction tools, Eurecom 3677, Avignon, France, 2012.

[10] 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.

[11] M. Röder, R. Usbeck, and A.-C. Ngonga Ngomo. Gerbil's new stunts: Semantic annotation benchmarking improved. Technical report, Leipzig University, 2016.

[12] 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.

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

[14] N. Steinmetz, M. Knuth, and H. Sack. Statistical analyses of named entity disambiguation benchmarks. In *Proc. of NLP & DBpedia 2013 workshop at 12th Int. Semantic Web Conference*. CEUR-WS, 2013.

[15] 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.

[16] R. Usbeck et al. GERBIL – general entity annotation benchmark framework. In *24th World Wide Web Conf.* ACM, 2015.

[17] 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 *10th edition of the Language Resources and Evaluation Conference*. ELRA, 2016.

[18] 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.* CEUR-WS, 2015.