All That Glitters is not Gold – Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking

Kunal Jha[1], Michael Röder[1], and Axel-Cyrille Ngonga Ngomo[1,2]

[1] AKSW Research Group
University of Leipzig
Augustusplatz 10, 04103 Leipzig, Germany
kunal.jha@uni-bonn.de,
roeder@informatik.uni-leipzig.de
[2] Data Science Group
University of Paderborn
Pohlweg 51, 33098 Paderborn, Germany
ngonga@upb.de

Abstract. The evaluation of Named Entity Recognition as well as Entity Linking systems is mostly based on manually created gold standards. However, the current gold standards have three main drawbacks. First, they do not share a common set of rules pertaining to what is to be marked and linked as an entity. Moreover, most of the gold standards have not been checked by other researchers after they were published. Hence, they commonly contain mistakes. Finally, many gold standards lack actuality as in most cases the reference knowledge bases used to link entities are refined over time while the gold standards are typically not updated to the newest version of the reference knowledge base. In this work, we analyze existing gold standards and derive a set of rules for annotating documents for named entity recognition and entity linking. We derive EAGLET, a tool that supports the semi-automatic checking of a gold standard based on these rules. A manual evaluation of EAGLET's results shows that it achieves an accuracy of up to 88% when detecting errors. We apply EAGLET to 13 English gold standards and detect 38,453 errors. An evaluation of 10 tools on a subset of these datasets shows a performance difference of up to 10% micro F-measure on average.

Keywords: Entity Recognition, Entity Linking, Benchmarks

1 Introduction

The number of information extraction systems has grown significantly over the past few years. This is partly due to the growing need to bridge the text-based document Web and the RDF[3]-based Web of Data. In particular, NER (Named Entity Recognition) frameworks aim to locate named entities in natural language documents while Entity Linking (EL) applications link the recognised entities to a given knowledge base (KB).

[3] Resource Description Framework, https://www.w3.org/RDF/

NER and EL tools are commonly evaluated using manually created gold standards (e.g., [13]), which are partly embedded in benchmarking frameworks (e.g., [20, 1]). While these gold standards have clearly spurred the development of ever better NER and EL systems, they have three main drawbacks: (1) They do not share a common set of rules pertaining to what is to be marked and linked as an entity. (2) Moreover, most of the gold standards have not been checked by other researchers after they have been published and hence commonly contain mistakes. (3) Finally, while in most cases the KB used to link the entities has been refined over time, the gold standards are typically not updated to the newest version of the KB.

We address this drawback of current NER/EL benchmarks through the following contributions: (1) We present a study of existing benchmarks that proposes a unified set of rules for creating NER/EL gold standards. (2) We present a taxonomy of common errors that can be found in the available gold standards that violate the rules. (3) We propose and evaluate EAGLET—a semi-automatic gold standard checking tool that is based on a fully automatic error detection pipeline. (4) We derive improved versions of 3 NER/EL benchmark subsets and quantify the effect of erroneous benchmarks on 10 NER/EL systems.

The rest of this paper is structured as follows. In the subsequent section, we give a brief overview of existing NER/EL gold standards. In Section 3, we define a set of rules for the annotation process and identify common annotation errors. EAGLET is described in Section 4 and evaluated in Section 5 along with state-of-the-art NER/EL tools on improved benchmarks. We conclude the paper with Section 6.

2 Related Work

While a large number of publications on new gold standards for the NER/EL tasks are available, only a few describe the process which led to their creation. In the following, we present a non-exhaustive list of English NER/EL benchmarks. *ACE2004* [17] was created using a subset of the ACE co-reference data set which was originally annotated with entities of the types person, organization, facility, location, geo-political entity, vehicle and weapon [3]. The annotations of the subset were linked to Wikipedia articles by Amazon Mechanical Turk workers with an inter-rater agreement of 85% [17]. *AIDA/CoNLL* [6] was created by annotating proper nouns in Reuters newswire articles. People, groups, artifacts and events were linked to the YAGO2 KB if a corresponding entity existed. *AQUAINT* [15] was created based on news articles. The documents were annotated automatically and checked manually. *DBpedia Spotlight*'s [13] evaluation dataset contains 60 natural language sentences from ten different documents with 249 annotated DBpedia entities overall. *IITB* [9] was created based on Web documents gathered from different domains. The authors explicitly state that emerging entities (EEs), i.e., entities that can be found in the text but are not present in the KB [7], should be annotated. *KORE50* [5] is a subset of the larger AIDA dataset. The selection of the KORE 50 dataset followed the objective to be difficult for disambiguation tasks. It contains a large number of first names referring to persons, whose identity needs to be deduced from the given context. However, the authors do not offer a list of the types of entities that have been annotated. *Microposts2014* [19] was created using a set of anonymized

twitter messages. Entities have been extracted using the NERD-Framework and linked to DBpedia articles manually by raters. After that, two experts double checked the ratings and managed conflicts. The dataset is separated into two parts—a training and a test dataset. *MSNBC* [2] was created based on news articles. An automatic NER and EL approach has been applied to generate the annotations following which have been checked manually. *OKE* [16] datasets have been created for the Open Knowledge Extraction Challenge 2015. 196 sentences have been annotated manually marking people, organizations, roles and locations.

Recently, Van Erp et al. [21] analyzed gold standards and concluded, that the available gold standards are diverse regarding several decisions that their creators have made during the creation process. However, their analysis focused primarily on the entities that have been marked and their characteristics instead of the correctness of gold standard annotations. Ehrmann et al. [4] presented a systematic overview of written and spoken natural language processing resources that can be used for named entity tasks like EL or NER. They pointed out that the quality of these resources is difficult to assess since many gold standards do not have a detailed documentation of the annotation process. In 2015, Ling et al. [11] presented a modular approach for the EL task which is motivated by the same observation as Van Erp et al., i.e., that a common understanding of the task is missing and several different interpretations are possible. The decisions that are made based on these interpretations have a huge impact on the design of a system and the gold standard that is used to benchmark this solution. Regarding the EL task, they list the following 5 major points for discussion:

P1) It is not defined whether only *named* entities or all resources in the given KB should be linked.
P2) It is not defined which entity should be chosen if more than one are plausible. The authors motivate this with the example of reoccurring events and different iterations of the same institution, e.g., the different United State Congresses. While these entities can be defined as not distinct to ease the problem, the authors argue that a statement like *"Joe Biden is the Senate President in the 113th United States Congress"* [11] can lead to wrong information if a system extracts Joe Biden as the President of all United States Congresses. On the other hand, they raise the problem that it might not be possible to formulate a statement about an event that will take place in the future since it might not be available in the KB.
P3) A similar problem is metonymy, i.e., an entity is called not by its own name but by another associated name. A common metonym for a government, e.g., the government of the United States, is the capital in which it is located, e.g., Washington. The authors write that linking to the capital entity as well as to the government entity is possible.
P4) There is no common set of entity types shared across different gold standards. For example, in some datasets, events are linked as entities while in other datasets, they are not.
P5) Following the authors, it is not clear whether annotations can overlap. In their example of an U.S. city which is followed by its state—*"Portland, Oregon"*—they argue to annotate the city, the state and both words together since all three markings make sense.

Rehm [18] defined a lifecycle for language resources. Our work can be used in the evaluation and quality control phase during the development of EL/NER gold standards to semi-automatically check the created corpus. Additionally, it can be used after the publication of the gold standards for its maintenance, i.e., to keep the gold standard up to date with new versions of the used KB.

3 Formal Annotation Framework

The creation of a NER/EL gold standard is a difficult task because human annotators commonly have different interpretations of this task as shown by [17]. It is, therefore, important to define a generic set of rules for annotating named entities in natural language text which leaves little if no room for interpretation. An advantage of having such rules is that they can be used to check gold standards automatically. The goal of this section is to present exactly such a set of rules derived from existing benchmarks. Based on the related work described in Section 2 we summarize assumptions that we can build upon. Thereafter, we define a set of rules for the preparation of a gold standard followed by a list of errors that we observed in existing gold standards.

3.1 Assumptions

We rely on the following assumptions:

A1) A single sentence does not need to have a linear structure. However, since state-of-the-art annotation systems do only annotate consecutive words, the gold standards should contain only annotations that can be expressed in this way. The word group *"Barack and Michelle Obama"* contains two persons. To annotate the first person, only the first name of Barack Obama can be annotated and linked to its entity. This assumption has the drawback that in the example *"Mr. and Mrs. Obama"* the word *"Mr."* would have to be linked to Barack Obama.

A2) The annotation should cover as many consecutive words as possible to represent the entity as precisely as possible. In the word group *"legendary cryptanalyst Alan Turing"* all these words should be part of a single annotation linked to the resource representing Alan Turing. However, this assumption should not be used to annotate whole clauses which will be described as Long Description Error later on.

A3) Each annotation should be linked to the most precise resource of the KB that is represented by the annotation or it should get a synthetically generated URI if this entity is an EE. Hence, in the example of point P2 described in Section 2, *"113th United States Congress"* has to be linked to a resource that represents exactly this 113th congress—not to the resource of the United States Congress in general.

A4) The annotated string should point to a specific entity. Indirect meanings of a string should not be considered. This assumption is important to make sure that a human annotator does not start to think laterally.

A5) The decision pertaining to which resources of a knowledge base can be used as entities for linking relies on a given set of entity types T_A. Only those entities that have at least one of the given types should be used for annotation.

3.2 Rule Set

Based on the aforementioned assumptions, we define a set of rules for marking the annotation of entities.

1. Consider each dataset D to be a set of documents and each document d to be an ordered set of words, $d = \{w_1, ..., w_n\}$.
2. Regard every word $w_i \in d$ as a sequence of characters or digits starting either at the beginning of the document or after a white space character and ending either at the end of the document or before a white space or punctuation character.
3. The annotation process relies on the set of entities $E = \{e|\tau(e) \cap T_A \neq \emptyset\}$ where τ is a function that returns the set of types T_e of the entity e and T_A is a given set of types that should be annotated in the corpus. It should be noted that E might contain more entities than the given KB K and that $E \backslash K$ is the set of EEs that can be found in the documents.
4. An annotation $a \in A$ is defined as $a = (S_a, u_a)$, where
   (a) S_a is a maximal sequence of consecutive words, such that $S_a = (w_i, w_{i+1}, w_{i+2}, \ldots)$ and
   (b) u_a is a URI that is used to link the annotated sequence to an entity $e = \delta(u_a)$, where δ is the dereferencing function returning the entity that can be identified with the given URI and e is
      i. the most precise entity possible
      ii. that represents a as described in A3.
5. The annotation function $\rho(d, K \cup E, T_A) = A$ creates a set of annotations $A = \{a_1, a_2, \ldots, a_n\}$ that meet the following requirements
   (a) $\delta(u_{a_i}) \in E$,
   (b) $\forall a_i, a_j \in A \left(S_{a_i}, S_{a_j} \subset d\right) \wedge \left(S_{a_i} \cap S_{a_j} = \emptyset\right)$ and
   (c) A has to be complete, i.e., it has to contain all valid annotations that can be found in d.

3.3 Comparison with Related Work

In this section, we compare our rules with the related work—especially with the points raised by Ling et al. [11] described in section 2. Rules 1 and 2 define the structure of a document and the words inside a document. Combined with rule 4.a, the possible positions of annotations are defined and the starting or ending of an annotation within a word is prohibited.

Rule 3 solves several issues that are raised in the related work. It answers **P1** by raising the requirement of a predefined entity type set on which the annotation process is based. A definition of the term *named entity* is not needed anymore and the exhaustive linking using all resources of the KB is only a special case in which the set of entity types comprises all types contained in the KB. It also solves **P4** by transforming the need of a common set of entity types that was bound to the unclear term *named entity* into a parameter of the annotation process.

Rule 4.b defines the linking step, i.e., the assignment of a URI to an annotated part of the text. With defining e as the most precise entity, the problem of the metonymy

described as **P3** is solved, since it becomes clear that *"Washington"* has to be linked to the U.S. government if it is used as its metonym. Note that the last part of the rule *"[...] that represents a directly"* does not object the linking of metonyms but prohibits the linkage of long descriptions which are described in the following section. It also prohibits the linkage of pronouns which aligns with our argumentation that pronouns should not be annotated since this would imply a NER/EL system to include a pronominal coreference resolution—an own, separated field of research that has lead to several solutions for this problem, e.g., the work of Lee et al. [10].

Together with the possible linking of EEs defined in rule 3, Rule 4.b solves **P2** as well. In cases in which a statement has to distinguish reoccurring events and different iterations of the same organization, these single events or organizations have to be linked to the most precise entity, i.e., one certain event or iteration. The argument, that events in the future could cause a problem is not valid since based on our rule set, this event would be handled as EE.

Rule 5 defines the annotation function that is based on the other 4 rules. Rule 5.b defines annotations as non-overlapping which answers the question raised in **P5**. According to rule 4.b, an annotation already contains the most precise link this particular part of the text could have. Adding additional annotations can lead to several problems. First, it would lead to a much larger amount of annotations without adding more information that couldn't be retrieved from the most precise entity, e.g., the fact that `dbr:Portland,_Oregon` is located in `dbr:Oregon`.[4] If this additional information is needed, it should be retrieved using available linked data technologies. Second, it can lead to an unnecessary shift of the focus, since the topic of the example is neither `dbr:Oregon` nor `dbr:Portland` but `dbr:Portland,_Oregon`.

3.4 Observations

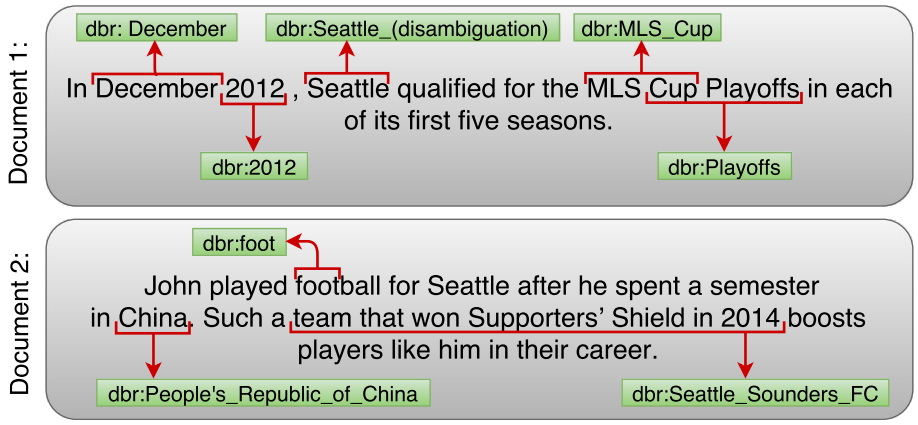

Fig. 1. Example documents.

[4] Throughout the paper, the prefix `dbr:` stands for `http://dbpedia.org/resource/`.

Having defined these generic rules, we evaluated the human annotated gold standards based on the aforementioned rules and assumptions. The evaluation unveiled various anomalies within the gold standards that we classified into the following categories.

Long Description Error (LDE). The first kind of error stands for annotations of sequences of words which might describe the entity they are linked to but do not contain a surface form of the entity (hence violating rule 4(b)ii).

For example, in Document 2 of Figure 1, *"a team that won Supporters' Shield in 2014"* is linked to `dbr:Seattle_Sounders_FC` but the marked text is neither equivalent to the surface form of entity nor directly describes the entity.

Positioning Error (PE). The next kind of error lies in marking a portion of a word in a sequence of words as an entity. Given that the rule 4a states that an annotation is only allowed to mark complete words, these errors violate rule 4a and the definition of words in rule 2. In Document 2 of Figure 1 for example, the *"foot"* in *"football"* is marked as an entity, hence violating the basic definition of the word.

Overlapping Error (OE). The third kind of error involves the presence of two or more annotations that share at least one word, thus violating the rule 5b. In Document 1, *"MLS Cup"* and *"Cup Playoffs"* have been marked over common part of the text *"Cup"*.

Combined Marking (CM). This is a non-trivial tier of errors wherein consecutive word sequences are marked as separate entities while the word sequences, if combined, can be annotated to a more specific entity. These errors are a direct violation of rule 4(b)i. In Document 1, *"December"* and *"2012"* are two separate consecutive entities which when combined together, *"December 2012"* are more apt in the context, i.e., link to the most precise resource.

URI Error. This error category comprises errors that violate rule 4b and can be separated into the following sub categories.

1. *Outdated URI (OU).* In this category, the entity is linked to an outdated resource which no longer exists in any KB. In Document 2, *"China"* is linked to `dbr:People's_Republic_of_China` which no longer exists in the KB but instead has to be updated to `dbr:China`.
2. *Disambiguation URI (DU).* This type of errors involves linking an entity to a non-precise resource page (disambiguation page) instead of a single resource. In Document 1, the entity *Seattle* is annotated with the URI `dbr:Seattle_disambiguation`, which is a disambiguation page that points to the City `dbr:Seattle` and the team `dbr:Seattle_Sounders_FC`. In this case, the team is the correct resource and should also be chosen as annotation.
3. *Invalid URI (IU).* This error category comprises annotations with no valid URI, e.g., an empty URI.

Inconsistent Marking (IM). This category comprises entities that were marked in at least one of the documents but whose occurrence in other documents of the same dataset is not marked as such. For example, the entity *Seattle* was marked in Document 1 but is left out in Document 2.

Missing Entity. The final categorisations of anomalies is a further extension of EM error. This comprises the presence of entities which satisfy the type conditions of the gold standard but were not been marked. This tier of error falls under the dataset completion and violates Rule 5c.

4 Eaglet

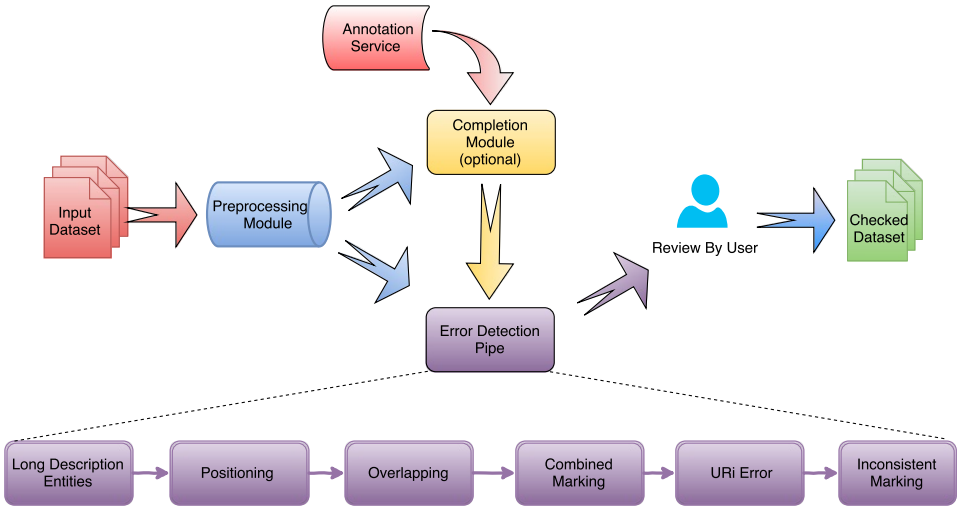

Fig. 2. Eaglet's Overview

The systematic classification of errors above allows for the creation of a framework, which can detect and correct a large portion of these errors. We hence present EAGLET (see Figure 2), a semi-automatic framework which aims at processing gold standards so as to detect the aforementioned anomalies and help rectify the errors, once reviewed by users.

4.1 Preprocessing Module

The input documents are first transformed into the structure described in Rule 1, i.e., each document is tokenized into an ordered set of words $d = \{w_1, ..., w_n\}$. Thereafter,

a POS-tagger and a lemmatizer are applied and the lemmas are attached to the words for later reusage.[5]

4.2 Completion Module

The completion module is an optional component. It uses publicly available annotation services to derive a list of entity annotations that are missing in the original dataset. These additional annotations support the work of a user that wants to make sure that the dataset is complete as defined in Rule 5c. However, since state-of-the-art annotation systems are not perfect [20], this module is based on a majority vote, i.e., the majority of the annotation systems have to contain an annotation inside their result list before it is added to the document.

For this module, we relied on the open-source project GERBIL that enables the usage of up to 13 different annotation systems [14].

4.3 Error Detection Pipeline

The error detection pipeline is the primary component of the tool. It tries to identify as many errors as possible in an automatic way based on the rules defined above. Every error type is handled by an own independent module enabling a particular configuration of the pipeline. Annotations that are identified as faulty are marked.

1. *Long Description Detection Module*: This module checks for the Long Description Error by searching for a relative clause inside an annotation.
2. *Wrong Positioning Detection Module*: This module searches for Positioning Errors by searching for mismatches between the start and ending positions of single annotations and the start and ends of words (see Rule 2).
3. *Overlapping Entity Detection Module*: This module checks for entity markings within each document whose positions are intersecting.
4. *Combined Tagging Detection Module*: This module searches for consecutive annotations that are separated by a white space character. Such entities are marked and a larger, combined annotation is generated and added to the document.
5. *URI Error Detection Module*: The URI checking module checks the URIs of all entities regarding their format. If a URI points to a reference KB, the module tries to dereference the URI to check whether a) the entity exists and b) the URI does not point to a disambiguation page. For example, if the given KB is the Wikipedia[6] or entities can be directly mapped to Wikipedia entities the module uses the Wikipedia API to determine whether the URI is outdated and derives the new URI.
6. *Inconsistent Marking Module*: This module collects all annotations in the corpus that have not been marked as faulty by one of the other modules. The lemmatized surface form of every annotation is used to search for occurrences of the entity inside the documents that have been missed. If such a surface form is identified, the module makes sure that the surface form can be marked following Rule 2 and

[5] We used the Stanford CoreNLP suite [12]

[6] `http://wikipedia.org`

that no annotation intersects with the identified occurrence before inserting a new annotation. Since these newly added annotations might be incorrect, e.g., because a URI that is linked to a word in one document does not need to fit to the same word in a different document, they are marked as added by the pipeline and should be checked by the user in the review module.

4.4 Review Module

The list of markings computed by the modules above is sent to the review module allowing the user to review the proposed changes in the dataset. The user interface of our tool allows every user to check each of the documents in the gold standards manually. Users can accept, modify or reject the suggestions of the tool as well as add new entities that have been missed by the completion module.

If a user adds a new entity annotation to a document, it is added to the completion module that processes the remaining documents again, searching for this new entity. This reprocessing aims at reducing the amount of entities the user has to add manually.

5 Evaluation

Our evaluation had three goals: First, we wanted to quantify the number of errors found in existing reference datasets.Secondly, we also wanted to know the accuracy with which EAGLET can detect errors. Finally, we aimed to quantify how much these errors in datasets influences the observed performance of NER/EL tools. We hence evaluated our approach within three different experiments.

5.1 Experiment I

In our first experiment, we ran EAGLET on the 13 datasets available in the GERBIL evaluation platform at the time of writing. The results are presented in Table 1 as a percentage of the total number of annotations (except for EM) found in each of the reference datasets. Our results show that errors of type PE, CM and URI errors occur often (e.g., up to 36% of CM errors in the IITB dataset) in all the datasets while the numbers for LDE are comparatively lower. OE were found only in DBpediaSpotlight, MSNBC, N3-Reuters-128, OKE2015 Task1 and IITB. We present absolute figures in case of Inconsistent Markings as it, unlike the other errors, involves adding entities to the list of existing annotations. Up to 9904 IM errors are found in a single dataset (IITB).

5.2 Experiment II

To evaluate the accuracy of EAGLET, we analysed a subset of the results of Experiment I manually. As pointed out in Section 2, only 4 datasets—ACE2004, AIDA/CoNLL and both OKE2015—come with a definition of the set of entity types that have been used for the annotation process. We randomly chose 25 documents from the ACE dataset, 25 documents from the AIDA/CoNLL dataset and 30 documents from the OKE evaluation

Table 1. Dataset features and amount of errors. (Abbreviations: $|D|$ = number of documents, $|A|$ = number of annotations, LDE = Long Description Error, PE = Positioning Error, OE = Overlapping Error, CM = Combined Marking, OU = Outdated URI, DU = Disambiguation URI, IU = Invalid URI, IM = Inconsistent Marking (in absolute numbers).)

Systems	Size		Percentage							IM
	$\|D\|$	$\|A\|$	LDE	PE	OE	CM	OU	DU	IU	
ACE2004	57	306	0.0	0.3	0.0	0.0	4.6	1.0	23.9	466
AQUAINT	1,393	34,929	<0.1	<0.1	0.0	4.0	2.0	0.2	12.2	6,357
AIDA/CoNLL-Compl.	50	727	0.0	0.0	0.0	8.4	10.3	1.4	5.8	586
DBpediaSpotlight	58	330	0.3	3.9	3.6	20.0	6.7	0.3	0.0	11
IITB	104	18,308	<0.1	1.8	0.3	36.0	4.5	7.7	<0.1	9,904
KORE50	50	144	0.0	3.4	0.0	0.0	11.1	0.0	0.0	3
Microposts2014-Test	1,055	1,256	0.2	2.1	0.0	5.8	3.2	0.3	0.4	698
Microposts2014-Train	2,340	3,822	0.2	2.3	0.0	6.6	2.8	<0.1	0.3	2,614
MSNBC	20	755	0.0	2.5	0.5	1.1	16.7	0.9	12.8	70
N3-RSS-500	500	1,000	0.0	0.1	0.0	0.0	2.4	0.0	0.1	193
N3-Reuters-128	128	880	0.3	0.8	0.2	1.6	4.1	1.5	0.9	111
OKE2015 Task1 eval	101	664	0.0	0.0	0.0	10.5	0.5	0.5	0.0	37
OKE2015 Task1 g.s.s.	96	338	0.0	0.0	1.2	6.8	2.4	0.0	0.0	52

Table 2. Results of the manual evaluation and the interrater agreement per task in brackets.

Dataset (subset)	Accuracy		Missed entities	
ACE2004	0.88	(0.89)	391	(0.81)
AIDA/CoNLL	0.80	(0.93)	71	(0.78)
OKE2015	0.79	(0.98)	14	(0.90)

dataset. Two researchers checked these documents independently, i.e., they evaluated the errors identified by the error detection pipeline. If at least one of them marked the pipelines decision for an annotation as wrong, the annotation was counted as error. Additionally, the two human annotators searched for entities that should have been marked according to the given type set but have been missed by the original gold standard creators and the error detection pipeline. Table 2 shows the accuracy of the error detection pipeline, the number of missed entities and the inter-rater agreement as F1-measure [8].

The automatic checking of the error detection pipeline was able to classify 79–88% of the annotations correctly. Especially the identification of URI errors worked well with an accuracy of 94%. The performance of the Combined Tagging Detection Module showed some minor flaws. For example, the name of the reporter of a news article directly followed by a city name, e.g., "Steve Pagani VIENNA" (AIDA/CoNLL dataset), was marked as two annotations that should be merged. The module should also be extended to deal with locations that are followed by the state in which they are located, e.g., "Grosse Pointe Park, Mich.". These annotations should be merged to fit the rule set and represent the entity, e.g., `dbr:Grosse_Pointe_Park,_Michigan`.

An important insight revealed by our evaluation is the large number of missed entities in current gold standards—especially for the ACE2004 dataset. The checked subset of the gold standard contained 190 annotations. The Inconsistent Marking Module added 14 correct annotations while the reviewers identified 195 additional annotations. 6 of the 25 documents did not contain any annotations at all in the original gold standard. *Not all that glitters is gold* and our results unveil that the ACE2004 gold standard is not really fit to be used for evaluating NER and EL systems.

We used the annotations added by the reviewers to evaluate the completion module. The module used the ten annotation systems listed in Figure 4. An annotation was counted as suggested if at least 5 systems marked it. It generated suggestions for 74%, 92% and 57% of the missing entities of the ACE, AIDA and OKE subsets, respectively. Note that these entities would have been considered mistakes as they did not exist in the reference data, clearly pointing towards the need for improved benchmarks.

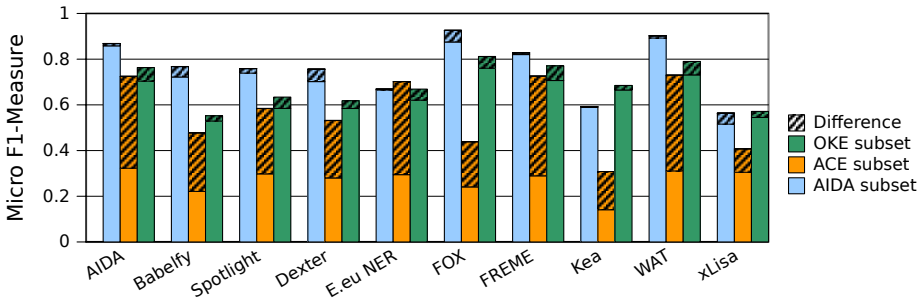

Fig. 3. NER Benchmark result differences of annotation systems on the original and corrected datasets.

5.3 Experiment III

The last experiment aimed to quantify the influence of the gold standard quality on the evaluation results of annotation systems. We used GERBIL to benchmark 10 annotation systems based on two versions of each of the three dataset subsets selected in Experiment II. The first version contained the original annotations while the second version was created based on the manual corrections of the output of the error detection pipeline. The annotation systems were tested using an A2KB (annotation to knowledge base) setting [20], i.e., the annotation systems received plain text, searched for named entities (NER) and linked them to the KB of the dataset (EL). Figure 4 shows the F1-scores for the original datasets as well as the difference to the F1-scores for the

corrected datasets.[7] Nearly all annotation systems[8] achieved a higher F1-score on the corrected subsets when compared with the original subsets. On average, the systems' F1-score increased by 16.4% for the ACE, 2.3% for the AIDA and was 1.5% higher for the OKE subset. The high influence of the gold standard quality on the benchmarking results can perhaps be seen most clearly in the ACE subset. While the *xLisa* annotator has a higher score than *DBpedia Spotlight*, *Dexter*, *Entityclassifier.eu NER* and *FREME NER* on the original subset, its performance is clearly lower on the corrected datasets.

To exclude the possibility that the results of the A2KB task were merely due to the EL step, we also computed the results of the frameworks on the NER substask. Our results (see Figure 3) show that the corrections have a high influence on the NER task as well. On average, the annotator performance increased with the correction of the ACE and OKE subsets by 29.3% and 4.4%, respectively. The highest enhancement with 43.7% and 6.4% was achieved by the *FREME NER* annotator. The average difference between the original and the corrected AIDA subset was 0%. While the performance of *Dexter* and *FOX* increased by 5.4% and 5.1% the F1-score of *xLisa* and *Babelfy* decreased by 4.8% and 4.6%.

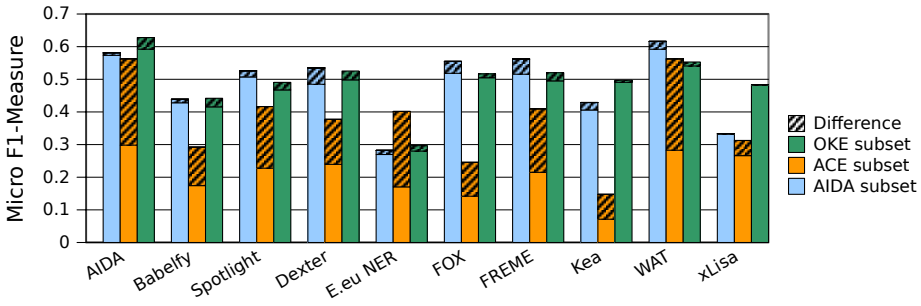

Fig. 4. NER and EL Benchmark result differences of annotation systems on the original and corrected datasets.

6 Conclusion

We derived a simple set of rules from common practice for benchmark creation. These rules were encoded into the benchmark curation tool EAGLET. A manual evaluation of EAGLET's results suggests it is a reliable tool for improving the quality of gold standards and thus improving the correctness of evaluation results for NER/EL tools. Within our evaluation of existing benchmarks, we were able to automatically detect a

[7] The complete result table can be found at `http://w3id.org/gerbil/experiment?id=201609290008`.

[8] The F1-scores of Entityclassifier.eu NER and xLisa for the corrected OKE subset were 1.7% and 0.2% percentage points lower than for the original subset.

significant amount of errors in a large number of corpora. The evaluation of the performance of systems on these datasets and the variation in their performance clearly underlines the importance of having gold standards which really achieve gold standard quality, i.e., which are free of errors. While we have noticed a move towards benchmarking platforms for NER and EL over the last years [20, 1], our results suggest the need for a move towards automatic benchmark checking frameworks, the first of which we provide herewith. However, they also suggest that alternative (if possible computer-assisted) approaches for the creation of benchmarks must be developed to ensure (1) the provision of benchmarks of high quality upon which (2) tools can be trained to achieve their best-possible performance. We hence regard this work as a first stepping stone in a larger agenda pertaining to improving the assessment of the performance of natural language processing approaches.

Acknowledgments

This work has been supported by the H2020 project HOBBIT (GA no. 688227) as well as the the EuroStars projects DIESEL (project no. 01QE1512C) and QAMEL (project no. 01QE1549C).

References

1. Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 249–260, New York, NY, USA, 2013. ACM.
2. Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716, 2007.
3. George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.
4. Maud Ehrmann, Damien Nouvel, and Sophie Rosset. Named entity resources - overview and outlook. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).
5. Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. KORE: keyphrase overlap relatedness for entity disambiguation. In *Proceedings of CIKM*, 2012.
6. Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, Michael Wiegand, and Gerhard Weikum. Robust disambiguation of named entities in text. In *proceedings of EMNLP 2011*, pages 782–792, Stroudsburg, PA, 27-31 July 2011. ACL.
7. Hoffart, Johannes and Altun, Yasemin and Weikum, Gerhard. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd WWW*, pages 385–396. ACM, 2014.
8. George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005.

9. Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD*, pages 457–466. ACM, 2009.
10. Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Conference on Natural Language Learning (CoNLL) Shared Task*, 2011.
11. Xiao Ling, Sameer Singh, and Daniel S Weld. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328, 2015.
12. Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.
13. Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian 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.
14. Röder Michael, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. Techreport for GERBIL 1.2.2 - V1. Technical report, Leipzig University, 2016.
15. David Milne and Ian H. Witten. Learning to link with wikipedia. In *17th ACM CIKM*, pages 509–518, 2008.
16. Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. *Open Knowledge Extraction Challenge*, pages 3–15. Springer International Publishing, 2015.
17. Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384. ACL, 2011.
18. Georg Rehm. The language resource life cycle: Towards a generic model for creating, maintaining, using and distributing language resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).
19. Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie, editors. *Proceedings, 4th Workshop on Making Sense of Microposts (#Microposts2014): Big things come in small packages, Seoul, Korea, 7th April 2014*, 2014.
20. Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*, 2015.
21. Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC 2016*, 05 2016.