Improving Named Entity Linking Corpora Quality

Albert Weichselbraun[1], **Adrian M.P. Braşoveanu**[2], **Philipp Kuntschik**[1], and **Lyndon J.B. Nixon**[2]

[1]Swiss Institute for Information Science, University of Applied Sciences, Chur, Switzerland
{*albert.weichselbraun,philipp.kuntschik*}*@htwchur.ch*
[2]MODUL Technology GmbH, Vienna, Austria
{*adrian.brasoveanu,lyndon.nixon*}*@modul.ac.at*

Abstract

Gold standard corpora and competitive evaluations play a key role in benchmarking named entity linking (NEL) performance and driving the development of more sophisticated NEL systems.

The quality of the used corpora and the used evaluation metrics are crucial in this process. We, therefore, assess the quality of three popular evaluation corpora, identifying four major issues which affect these gold standards: (i) the use of different annotation styles, (ii) incorrect and missing annotations, (iii) Knowledge Base evolution, (iv) and differences in annotating co-occurrences. This paper addresses these issues by formalizing NEL annotations and corpus versioning which allows standardizing corpus creation, supports corpus evolution, and paves the way for the use of lenses to automatically transform between different corpus configurations. In addition, the use of clearly defined scoring rules and evaluation metrics ensures a better comparability of evaluation results.

1 Introduction

Named Entity Linking (NEL) systems identify mentions of named entities and link them to Knowledge Bases (KBs) (Ji et al., 2017). Their evaluation heavily depends upon annotated gold standards and competitions such as TAC-KBP (Ji et al., 2017) or Open Knowledge Extraction (Nuzzolese et al., 2016) which help drive research and advance the state of the art. Knowledge Bases, annotation guidelines and gold standards, NEL tools, as well as the evaluation systems themselves, were found to introduce errors into NEL evaluations (Braşoveanu et al., 2018). The most critical issues are related to corpora quality due to wrong, partial or insufficient annotations (van Erp et al., 2016; Jha et al., 2017). Annotation guidelines used to produce a gold standard come with different rules for describing whether the annotation systems should take into account overlapping mentions, co-references or general concepts mentioned in a KB (Rosales-Méndez et al., 2018). These guidelines are domain-specific and often depend on the application and task-specific context. Semantic search, for instance, would not require co-references, but they are relevant for relation extraction tasks (Rosales-Méndez et al., 2018).

The mentioned issues can be approached from various angles. General improvements such as a clear definition for NEL will affect multiple components of a NEL system. More specific suggestions, such as KB refinements typically affect only corpora that link entities to that particular KB. In all these cases the objective should be improving the quality and transparency of both the corpus and the evaluation processes.

A serious issue with current NEL evaluations is the lack of flexibility during the evaluation process which forces NEL systems to adapt to the used evaluation corpora. For example, if a KB has more name variants (e.g., Bobby Kennedy and RFK for Robert F. Kennedy) than the corpus annotators have considered, NEL systems able to correctly detect these name variants will be penalized since they do not occur in the corpus and are, therefore, considered errors. This paper, therefore, suggests solutions such as lenses (Section 3) and corpus versioning (Section 4.2) to address this issue. In particular, our main contribution is describing the innovations that can be applied at a corpus and evaluation systems level in order to create more flexible and expressive evaluations.

The rest of the paper is organized as follows: Section 2 describes related work; Section 3 collects empirical evidence for common issues with NEL corpora and provides a concise problem description; Section 4 presents our approach towards building flexible evaluation systems; whereas the final section discusses the conclusions.

2 Related Work

An early analysis on the effect entity overlap between different data sets, confusability (the number of meanings a surface form can take) and dominance in several data sets was performed in (van Erp et al., 2016). Most current gold standards are known to suffer from *annotation mistakes*, *lack of updates* (typically due to the fact that there are no clear updating guidelines or funds for this operation), *popularity bias* (tools return most popular candidates), *small volume* (only several hundred examples), and are typically focused on a small set of languages (Ngomo et al., 2018).

Early methods to improve the quality and speed of human annotations have included: dynamic sentence selection in combination with iterative pre-annotation and qualitative checks (Tsuruoka et al., 2008) or crowdsourcing (Sabou et al., 2014); whereas more recent techniques include hypergraphs in order to highlight various name variants (Katiyar and Cardie, 2018). Several recent linguistic-driven techniques for improving gold standard quality discussed in (Sakor et al., 2018) include effect of capitalization, of implicit/explicit entities, of number of words or hidden relations. Automatically generated benchmarks like those provided with BENGAL (Ngomo et al., 2018) can help improve annotation speed, but only if deployed together with advanced debugging and error analysis tools, as otherwise there will be a risk of increasing bias.

A set of KB improvements can also be devised to aid domain experts and NEL systems in identifying mentions of named entities. KBs, for example, can be used to expand upon the number of name variants from a corpus by including multiple annotations for each entity to cover cases of embeddings, overlap and extensions (Odoni et al., 2018). At the named entity disambiguation level, such techniques include collecting all name variants from multiple Knowledge Graphs (Ehrmann et al., 2016) or using hypergraphs and multi-layer bi-LSTMS to detect the nested entities (Katiyar and Cardie, 2018).

The lack of parallel language corpora like Europarl (Koehn, 2005) for NEL (except for some smaller corpora like MeanTime (Minard et al., 2016)) is another serious issue that is rarely discussed, though such corpora are available for NER (Agerri et al., 2018).

3 Lenses and Corpus Quality

Today's NEL evaluation tools automatically compare the performance of multiple NEL systems on various data sets using a wide variety of experiments (e.g., NER - Named Entity Recognition, Entity Typing, D2KB - matching entity mentions to certain KB, etc) and indicators (e.g., precision, recall, F1, accuracy, etc). GERBIL (Röder et al., 2018) and its extensions (Waitelonis et al., 2019) were designed to support multiple experiment types using black box evaluation techniques. The neleval system (Hachey et al., 2014) based on the TAC-KBP guidelines provides primary error explanations. As an alternative, visual evaluation systems such as Orbis (Odoni et al., 2018) and VEX (Heinzerling and Strube, 2015), also allow close inspection of the evaluation results and help designers improve system performance. Upon analyzing these packages, we came to the conclusion that none of them provides suitable tools for handling multiple annotation styles and updating gold standards. Therefore, reuse of old gold standards can lead to problematic results (e.g., entities declared NIL in the gold can currently exist in the current KB version and can be retrieved by annotator tools) or even unfair evaluations (e.g., tools that use an old KB should not be compared with those who use the latest updates).

We think NEL systems should be evaluated against both updated gold standards and KBs, therefore we suggest the application of lenses over the existing data, i.e. transformations between different KB versions (e.g., DBpedia 2015-10 and DBpedia 2017-10), KBs (e.g., DBpedia and Wikidata) and annotation styles (e.g., always take the longest strings or only annotate non-NIL entities). The following section presents an analysis of three popular NEL corpora, discusses different use case for lenses and the corresponding transformation rules that are required to transfer corpora from one representation to the other.

3.1 Corpus Quality

Creating high quality gold standard data is a difficult task due to: (i) *incomplete annotation rules* - i.e. cases that have not been properly covered in the annotation rules; (ii) *errors* - present on multiple levels, from human or automated annotation errors, to process errors (e.g., errors in the annotation guidelines) or KB errors; as well as (iii) *decay* - KBs used for the annotation might become obsolete or outdated, forcing the corpus maintainer to consider KB evolution or even KB migration.

In order to understand the extent to which a corpus may require updating, we have analyzed several well-known corpora from our field: (i) *OKE2016 Task 1* (Nuzzolese et al., 2016) - focuses on short biographical sentences extracted from Wikipedia; (ii) *Reuters128* (Röder et al., 2014) - a set of texts extracted from the classic Reuters corpora; and (iii) the English partition of the *MeanTime* (Minard et al., 2016) corpus that covers events extraction.

Table 1 illustrates several attributes related to the number of entities in a corpus that directly reflect *incomplete annotation rules*, *errors* and *decay*: (i) the number of original annotations; (ii) the number of NIL annotations at publication date; (iii) updated NIL counts which indicate the impact of KB evolution; (iv) count of potential overlaps within the original annotations that need to be handled according to the used annotation guidelines; (v) potential missing entities based on the annotation the guideline (if such a document exists).

Empty columns (marked with -) were not filled due to lack of available data or guideline. Counts from rows (i) and (ii) were taken directly from the corpora; row (iii) count was estimated based on SPARQL queries that aim at linking NIL entities to the KB; and counts for columns (iv) and (v) were estimated based on annotating samples from each data set.

Two annotators have independently annotated a quarter of the documents that have been selected using random sampling. Only the entities mentioned in the article or guideline related to a data set were used for the respective counts (e.g., Reuters128 was only annotated with Person (PER), Organization (ORG) and Location (LOC) based on (Röder et al., 2014)). The analysis only considers NIL entities that were marked as such (e.g., NIL or similar designation). Consequently, the OKE 2016 counts shows no NIL entities, since they were not included in the original dataset. The updated NIL counts are obtained by using SPARQL queries that determine whether NIL entities have become available in new KB versions.

Table 2 examples were extracted from the Reuters128 corpus and illustrate (i) missing and wrong groundings, (ii) KB evolution, and (iii) surface forms deviations due to different annotation rules.

Overlaps tend to appear in cases related to LOC and ORG entities. Quite often, an overlap is identified in long names such as *Chattanooga State Technical College* or *City University of New York Graduate Center*. In similar cases a surface form expansion that will contain the longest possible string should correctly match the entity from the gold standard. Complicated cases like the following: *Loyola's University in Belgium, Economics* (from OKE2016) can be interpreted in multiple ways. This example can either be rendered as (i) one long entity that corresponds to the whole string; (ii) one entity that describes the University (*Loyola's University in Belgium*); (iii) two entities (*Loyola's University* and Belgium); (iv) two entities again (*Loyola's University in Belgium* and a string *Economics*); even (v) three entities (*Loyola's University* and *Belgium* and *Economics*). The phrasing of the examined sentence suggests that the fourth version is the correct one. Without a thorough text analysis such instances are extremely difficult to disambiguate for both humans and machines.

Even if we leave aside difficult cases that typically show a low inter-rater agreement such as *Potential Overlap* (Table 1, row iv; e.g., agreement of 0.40 for Reuters128) or *confusability* (van Erp et al., 2016), there are still many mentions that are not spotted in the original corpus such as those from the *Missing Entities Guideline* (Table 1, row v) for which experts also exhibit better inter-rater agreements (e.g., 0.61 for Reuters128).

These findings suggest that the methods and processes used for annotating documents need to be updated. Applying lenses would be one of the methods that could address some of the mentioned shortcomings, as they would help both accounting for multiple points of views when annotating, as well as for KB evolution. Such lenses coupled with well-defined metrics for measuring NEL performance are key towards reliably assessing a system's performance and driving its development.

No	Rule	OKE2016	Reuters128	MeanTime
i	Original Annotation ALL	176	880	853
ii	Original NIL Only	-	230	554
iii	Updated NIL Count	-	175	465
iv	Potential Overlap	84	104	221
v	Missing Entities Guideline	76	180	272

Table 1: Estimated entity counts based on different criteria in three corpora. All data sets are in English.

surface	gold link	correct link	error
[Volkswagen AG] **[VOWG.F]**, [VW], is due ...	NIL	dbr:Volkswagen	Missing Annotation
bid for **[Avondale Mills]** ...	NIL	dbr:Avondale_Mills	KB evolution
[The Chicago Mercantile Exchange], **[CME]**, said ...	dbr:CME_Group	dbr:Chicago _Mercantile_Exchange	Incorrect Link
... of **[Salem**, Ore.]	dbr:Salem,_Oregon	dbr:Salem,_Oregon	Different surface form

Table 2: Examples of dataset errors. Gold entity spans are marked by parentheses. Errors are presented in bold. Abbreviations or tickers should be separate entities. Locations can include states.

3.2 Context-Specific and Application-Specific Transformation Rules for Lenses

A named entity linking system links a mention $m_{[s_i]}^{e_i,KB}$ or $m_{[x_i,y_i]}^{e_i,KB}$ of a named entity with surface form s_i within a document d to the corresponding entity e_i in a knowledge base KB. The variable x_i indicates the mention's start position within the document and y_i the corresponding end position.

Mentions may overlap and the specification of the knowledge base KB can be omitted, if it is not relevant for the application (e.g. if we do not consider different KB versions in the given use case).

The system distinguishes between

1. $m_{[s_i]}^{e_i,KB}$ surface forms s_i that were linked to an entity e_i within a knowledge base KB,
2. $m_{[s_i]}^{nil}$ mentions of Named Entities (NEs) that are not available in the KB and, therefore, are not linked (i.e. NIL entities), and
3. $m_{[s_i]}^{\emptyset}$ candidate mentions with surface form s_i that do not refer to a named entity.

3.2.1 Different Annotation Styles

Annotation styles specify rules that aid annotators in assessing if (i) a candidate mention should be considered a mention of a named entity, and (ii) the extent of the corresponding surface form.

Although a trivial design decision for isolated mentions, the consistent handling of nested mentions requires more thought. For instance, the text snippet *University of Western Australia Cricket Club* may contain, dependent on the applied annotation rule, up to four overlapping mentions (*Australia*, *Western Australia*, *University of Western Australia*, *University of Western Australia Cricket Club*). In addition, annotation styles might be entity type specific even within a single corpus.

We consider the following three annotation styles, as illustrated based on the annotation of the text snippet *Vienna, VA*:

1. *ØMIN* disregards overlapping entities and tries to extract the minimum number of entities: $m_{\text{[Vienna, VA]}}^{dbr:Vienna,\_Virginia}$, i.e. links the snippet to the *Vienna, Virginia* DBpedia entity.
2. The annotation style *ØMAX*, in contrast, extracts the maximum number of entities from a given text snippet: $m_{\text{[Vienna]}}^{dbr:Vienna,\_Virginia}, m_{\text{[VA]}}^{dbr:Virginia}$
3. The style *OMAX* allows for overlaps and, again, extracts the maximum number of entities: $m_{\text{[Vienna, VA]}}^{dbr:Vienna,\_Virginia}, m_{\text{[VA]}}^{dbr:Virginia}$

The presented rules only consider borderline cases, even though combinations of them can also be used within a corpus. For instance, a corpus might use the *OMAX* rule for LOC entities but apply *ØMIN* for all other entity types, therefore only yielding $m_{\text{[ETH Zurich]}}^{dbr:ETH\_Zurich}$ rather than

$m^{dbr:ETH\_Zurich}_{\text{[ETH Zurich]}}$ and $m^{dbr:Zurich}_{\text{[Zurich]}}$ for the text snippet *ETH Zurich*.

Table 3 outlines transformation rules between different annotation styles.

3.2.2 Knowledge Base Evolution

Lenses are also able to capture KB evolution, i.e. the case where a KB evolves due to changes or extended coverage of the underlying domain. Changes to the KB may

1. introduce new entities (e.g. the company Alphabet Inc. in October 2015),
2. lead to the deletion of entities that are no longer considered relevant, or
3. drive the introduction of a more fine grained or coarser mapping for existing entities.

Table 4 introduces the corresponding transformation rules. Newly introduced entities may enable the grounding of *NIL* entities to the extended knowledge base. The removal of an entity, in contrast, may transform an existing grounding to a *NIL* entity since the corresponding KB entity is no longer available. Finally, changes in granularity may either lead to the introduction of additional entities, or to the deletion of links to the KB.

3.2.3 Knowledge Base Migration

KB migration is the case in which a corpus that has been initially annotated with one KB is used to evaluate a component that links mentions to another KB. Many well maintained knowledge bases such as DBpedia, GeoNames and Wikidata contain links to indicate equivalent entities (e.g., *owl:sameAs*, *skos:exactMatch*, etc). These links and techniques such as ontology alignment may be used to automatize the transformation of a mentions $m^{e_i,KB}_{[x_i,y_i]}$ within a KB (KB) to the corresponding mention $m^{e_i,KB'}_{[x_i,y_i]}$ in the target KB (KB'). KB migration draws at the same set of transformation rules as the KB evolution use case.

3.2.4 Co-references

Co-references play a crucial role in natural language processing tasks such as relation extraction. A co-reference is a mention with surface form s'_i that refers to the same entities e_{ij} as other mentions $m^{e_{ij}}_{[s_{ij}]}$ within the document. Its surface form often contains prepositions or noun-phrases that on their own do not provide enough context to determine the referred entities e_{ij}. For anaphora and cataphora the co-reference $m^{e_i,KB}_{[s_i]}$ points to a single entity, for split antecedents the co-reference refers to multiple NEs. For instance, in the text: *"Berlin, Rome and Paris are capitals of European countries. These cities are also popular tourist destinations."* the surface form *These cities* refers to three previous named entities and is, therefore, annotated as $m^{dbr:Berlin}_{\text{[These cities]}}$, $m^{dbr:Rome}_{\text{[These cities]}}$ and $m^{dbr:Paris}_{\text{[These cities]}}$. Systems and corpora that do not support co-reference resolution, therefore, consider co-references as candidate mentions $m^{\emptyset}_{[s_i]}$ that do not link to any named entity.

3.3 Limitations of Corpus Transformation with Lenses

The rules outlined in the previous section can be used to translate between different corpus representations. Translations from expressive representations to less expressive ones can be done automatically and exposed to users as lenses. For example, a transformation from the OMAX to the ØMIN annotation style, from a corpus with annotated co-references to a corpus which does not considers them, and the KB migration use case for NIL entities may be performed automatically.

Otherwise, corpus versioning (Section 4.2) is required to record any changes added by manual or semi-automatic processes.

4 Method

This section discusses options for improving corpus quality by (i) introducing semi-automatic tools that support corpus creation and evaluation by automatically spotting violations of the annotation style and suspicious entities (Section 4.1), and (ii) suggesting guidelines for versioning corpora ensuring that improvements and extensions are incorporated in a meaningful and backward compatible way (Section 4.2).

4.1 Corpus Analysis Tools

Software developers frequently use static code analysis tools such as pylint[1], findbugs[2] and checkstyle[3] as part of the build pipeline to enforce coding style guidelines and to spot potential bugs in an early stage.

[1] www.pylint.org
[2] findbugs.sourceforge.net
[3] checkstyle.sourceforge.net/

Table 3: Lense transformation rules between different annotation styles.

Annotation style	ØMIN	ØMAX	OMAX
Corpus entity	$m^{e_1,KB}_{[x1,y1]}$	$m^{e_1,KB}_{[x1,y11]}, \ldots, m^{e_n,KB}_{[x1n,y1]}$	$m^{e_1,KB}_{[x1,y1]}, \ldots, m^{e_n,KB}_{[x1,y1]}$
Transformation to ØMIN	$m^{e_1,KB}_{[x1,y1]}$	$m^{e_1,KB}_{[x1,y1]}$	$m^{e_1,KB}_{[x1,y1]}$
ØMAX	$m^{e_1,KB}_{[x1,y11]}, \ldots, m^{e_n,KB}_{[x1n,y1]}$	$m^{e_1,KB}_{[x1,y11]}, \ldots, m^{e_n,KB}_{[x1n,y1]}$	$m^{e_1,KB}_{[x1,y11]}, \ldots, m^{e_n,KB}_{[x1n,y1]}$
OMAX	$m^{e_1,KB}_{[x1,y1]}, \ldots, m^{e_n,KB}_{[x1,y1]}$	$m^{e_1,KB}_{[x1,y1]}, \ldots, m^{e_n,KB}_{[x1,y1]}$	$m^{e_1,KB}_{[x1,y1]}, \ldots, m^{e_n,KB}_{[x1,y1]}$

Table 4: Lense transformation rules for knowledge base evolution and knowledge base migration.

Task	new entity	deleted entity	more fine grained entity mapping	coarser entity mapping
Corpus entity	$m^{nil,KB}_{[x_i,y_i]}$	$m^{e_i,KB}_{[x_i,y_i]}$	$m^{e_i,KB}_{[x_i,y_i]}$	$m^{e_{i1}}_{[x_{i1},y_{i1}]}, \ldots, m^{e_{in},KB}_{[x_{in},y_{in}]}$
Transformation	$m^{e_i,KB'}_{[x_i,y_i]}$	$m^{nil,KB'}_{[x_i,y_i]}$	$m^{e_{i1},KB'}_{[x_{i1},y_{i1}]}, \ldots, m^{e_{in},KB'}_{[x_{in},y_{in}]}$	$m^{e_i,KB'}_{[x_i,y_i]}$

We strongly believe that similar tools could be highly beneficial for aiding researchers in the creation and validation of NLP corpora, by

1. automatically locating violations of annotation styles (e.g. overlaps in case of a non-overlapping annotation style).

2. drawing upon POS tagging and dependency parsing for marking unusual annotations such as NEs that do not contain a noun to flag potentially incorrect annotations.

4.2 Corpus Versioning

Even corpora that are frequently used in NEL evaluation suffer from quality issues (see Table 1). Although addressing these issues is important, backward compatibility of refined corpora is key to their usefulness since it ensures that results can be compared to previously published work.

We, therefore, suggest corpus versioning to promote the improvement of corpora. Publishing multiple corpus versions will enable researchers to run evaluations against these versions and, therefore, provides means to compare the gathered results to other work.

Corpus versioning is needed for cases where an automatic translation to the desired gold standard representation via lenses is not feasible:

- addressing data quality issues and mistakes in the original corpus,

- the linking of *NIL* entities to a knowledge base entity due to knowledge base evolution or knowledge base migration, and

- a new more expressive annotation style.

Versioning should enable researchers to address these issues while ensuring

- a clear relation to the original corpus that makes comparison with previous versions feasible;

- support for multiple versions and version trees that have been contributed by different people and organizations (Figure 1)

- that corpus metadata provides (i) information on the relations between different corpus versions, (ii) easily traceable contributions, (iii) credits for contributors, and (iv) easy re-use of refined corpora for further evaluations;

- easily accessible corpus versions, e.g. by uploading them to research data platforms and providing a digital object identifier (DOI) for each version, so that researchers can easily cite, locate and re-use corpora.

Table 5: Lense transformation rules for co-reference resolution.

Task	single co-reference	split antecedents
Corpus entity	$m^{e_i}_{[s'_i]}$	$m_{[s'_i]}{}^{e_{i1}}, \ldots m_{[s'_i]}{}^{e_{in}}$
No co-reference resolution	$m^{\emptyset}_{[s'_i]}$	$m^{\emptyset}_{[s'_i]}$

Table 6: Suggested corpus metadata

Metadata	Description	Example
corpus_name	A name that identifies the corpus.	OKE2018
corpus_url	The corpus archive URL.	http://github.com/fhgr/oke2016-dbpedia-2019-02-01_v1.zip
creator	A comma-separated list of persons or organizations that created the corpus.	Sue May <sue@myorg.edu>
date	The corpus's publishing date.	2019-05-30
description	A description of the current corpus version.	Adapted OKE2016 to DBpedia 2019-02-01 and integrated bug fixes from 2018-03-07.
final	Is it usable for official evaluations?	false
parent_corpus_url	The URL of the parent corpus (if any)	http://github.com/fhgr/oke2016-dbpedia-2018-09-03_v2.zip
considers_corpus_url	A list of URLs pointing to related corpus versions.	http://github.com/sue/oke2016-dbpedia-fixes-2018-03-07.zip
annotation_style	A list of annotation styles per supported named entity type.	PER: NOMIN, GEO: OMAX, ORG: OMIN
annotators_per_document	Number of annotators per document.	3
annotator_agreement	Inter-rater-agreement between annotators computed using the Fleiss' kappa.	0.61

The key here is making evaluations easier to replicate and increase the benefit they provide to the community. Currently evaluations are often tightly designed to a specific context such as a competition or an application domain. These kinds of results are helpful for determining the best performing system under tight restrictions, but they unnecessarily restrict the scope of the evaluation. For instance, such results do not provide information on how systems cope with different annotation rules, settings and use cases. Automatically performing evaluations with all available lenses, corpus versions and scoring rules (Section 4.3) could address this issue. Scientists could still publish the results for their particular application context in the research paper but would in addition provide a DOI to the full results that cover also settings for which their system hasn't been optimized. Ultimately such an approach would improve the usefulness of evaluations since it would provide (i) much more context on the strengths and weaknesses of NEL systems, and (ii) broader insights into the effects of the suggested methods and design decisions.

4.2.1 Publishing a Corpus Version

We recommend a standardized directory structure for publishing corpora that contains:

1. a `corpus` directory containing all corpus data and annotations in the NIF format.
2. a `METADATA.yaml` file that describes the corpus based on the metadata introduced in Table 6.
3. a `README.md` file which provides additional unstructured information.

Popular version control services such as github and gitlab offer a release feature that automatically

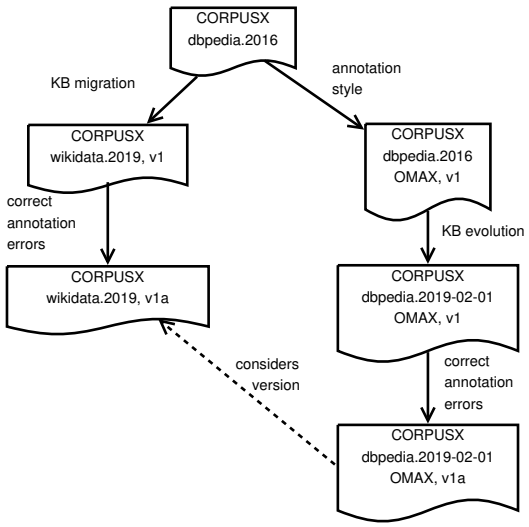

Figure 1: Common corpus versioning use cases.

publishes an archive of the released repository version which can be used to publish a certain corpus version. Another option would be publishing corpus versions in research data repositories such as zenodo.org which also provide a DOI and bibliographical metadata to data artifacts.

4.3 Scoring Rules

Scoring rules outline the conditions under which a gold standard corpus mention $m^c := m^{e_c,KB}_{[x_c,y_c]}$ and a mention returned by the NEL system $m^s := m^{e_s,KB}_{[x_s,y_s]}$ are considered equivalent to each other. The following three scoring rules are frequently used in NEL evaluations:

1. perfect match $\mathcal{P}$ - the entities refer to the same KB entity e_i, and the exactly same surface form s_i.

2. contained match $\mathcal{C}$ - both entities refer to the same KB entity e_i and the surface form of the mention returned by the NEL system m^s is contained in the surface form of the corpus mention m^c, i.e. $x^s_i \geq x^c_i$ and $y^s_i \leq y^c_i$.

3. overlapping match $\mathcal{O}$ - this case is equivalent to the contained match but further relaxes the restrictions on the surface form, so that even an overlap (i.e. $y^s_i \geq x^c_i$ and $x^s_i \leq y^c_i$) between entities is considered a valid match.

The used scoring rule have a significant impact on the computation of the NEL system's performance metrics such as precision and recall.

5 Conclusion

This paper discusses approaches for addressing the issues of corpus quality and the comparability of evaluations that have been performed with these corpora, together with associated annotations[4]. We discuss (i) factors that seriously affect the accuracy of evaluation corpora such as different annotation styles, missing and wrong annotations, KB evolution and co-reference handling.

In addition we also shed light on the issue of KB migration, which is relevant if the evaluation corpus and the NEL system use different KBs. Afterwards we (ii) introduce a formalization that captures these factors, and (iii) present transformation rules between different corpus configurations.

These transformation rules that expose different corpus configurations as lenses in conjunction with corpus analysis tools and corpus versioning are key towards improving corpus quality. Well-defined scoring rules and evaluation metrics are further steps towards standardizing evaluations and improving their validity and reproducibility.

Future research will focus on (i) applying these guidelines to the NEL evaluation of annotation of TV-related content in the ReTV project[5] so that results can be compared and evolved in the future if need be,(ii) the creation of tools that support corpus creation and evaluation processes, (iii) adding support for corpus versioning and the parallel analysis of multiple corpus versions to evaluation tools such as Orbis (Odoni et al., 2018) and GERBIL (Röder et al., 2018), and (iii) proving a research data infrastructure for publishing evaluation corpora and evaluations that have been performed on these corpora.

Acknowledgments

This work has been supported by the MedMon project (https://www.htwchur.ch/medmon) funded by Innosuisse and the ReTV project (https://retv-project.eu/) funded through the European Union's Horizon 2020 Research and Innovation Programme under GA No 780656 and through the FFG project EPOCH (https://epoch-project.eu).

[4] Annotations discussed in Section 3 are available at https://github.com/orbis-eval/corpus_quality_paper

[5] To be published in D1.2 deliverable at `https://retv-project.eu/deliverables/`

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*. European Language Resources Association (ELRA). http://www.lrec-conf.org/lrec2018.

Adrian M. P. Braşoveanu, Giuseppe Rizzo, Philipp Kuntschick, Albert Weichselbraun, and Lyndon J.B. Nixon. 2018. Framing named entity linking error types. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France, pages 266–271. http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html.

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors. 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA). http://www.lrec-conf.org/lrec2016.

Maud Ehrmann, Damien Nouvel, and Sophie Rosset. 2016. Named entity resources - overview and outlook. In (Calzolari et al., 2016). http://www.lrec-conf.org/proceedings/lrec2016/summaries/987.html.

Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*. The Association for Computer Linguistics, pages 464–469. http://aclweb.org/anthology/P/P14/P14-2076.pdf.

Benjamin Heinzerling and Michael Strube. 2015. Visual error analysis for entity linking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, System Demonstrations*. ACL, pages 37–42. http://aclweb.org/anthology/P/P15/P15-4007.pdf.

Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. 2017. All that glitters is not gold - rule-based curation of reference datasets for named entity recognition and entity linking. In Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and Olaf Hartig, editors, *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*. volume 10249 of *Lecture Notes in Computer Science*, pages 305–320. https://doi.org/10.1007/978-3-319-58068-5_19.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, and Cash Costello. 2017. Overview of TAC-KBP2017 13 languages entity discovery and linking. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST. https://tac.nist.gov/publications/2017/papers.html.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Association for Computational Linguistics, pages 861–871. https://aclanthology.info/papers/N18-1079/n18-1079.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. volume 5, pages 79–86.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In (Calzolari et al., 2016). http://www.lrec-conf.org/proceedings/lrec2016/summaries/488.html.

Axel-Cyrille Ngonga Ngomo, Michael Röder, Diego Moussallem, Ricardo Usbeck, and René Speck. 2018. BENGAL: an automatic benchmark generator for entity recognition and linking. In Emiel Krahmer, Albert Gatt, and Martijn Goudbeek, editors, *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*. Association for Computational Linguistics, pages 339–349. https://aclanthology.info/papers/W18-6541/w18-6541.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Robert Meusel, and Heiko Paulheim. 2016. The second open knowledge extraction challenge. In Harald Sack, Stefan Dietze, Anna Tordai, and Christoph Lange, editors, *Semantic Web Challenges - Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*. Springer, volume 641 of *Communications*

in Computer and Information Science, pages 3–16. https://doi.org/10.1007/978-3-319-46565-4_1.

Fabian Odoni, Philipp Kuntschik, Adrian M. P. Braşoveanu, and Albert Weichselbraun. 2018. On the importance of drill-down analysis for assessing gold standards and named entity linking performance. In Anna Fensel, Victor de Boer, Tassilo Pellegrini, Elmar Kiesling, Bernhard Haslhofer, Laura Hollink, and Alexander Schindler, editors, *Proceedings of the 14th International Conference on Semantic Systems, SEMANTICS 2018, Vienna, Austria, September 10-13, 2018*. Elsevier, volume 137 of *Procedia Computer Science*, pages 33–42. https://doi.org/10.1016/j.procs.2018.09.004.

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

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL - benchmarking named entity recognition and linking consistently. *Semantic Web* 9(5):605–625. https://doi.org/10.3233/SW-170286.

Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What should entity linking link? In Dan Olteanu and Barbara Poblete, editors, *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, Cali, Colombia, May 21-25, 2018.*. CEUR-WS.org, volume 2100 of *CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2100/paper10.pdf.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 859–866. http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.

Ahmad Sakor, Saeedeh Shekarpour, Maria-Esther Vidal, Jens Lehmann, and Sören Auer. 2018. Old is gold: Linguistic driven approach for entity and relation linking of short text pages 339–349.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2008. Accelerating the annotation of sparse named entities by dynamic sentence selection. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, Columbus, Ohio, pages 30–37. https://www.aclweb.org/anthology/W08-0605.

Marieke van Erp, Pablo N. Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*. pages 4373–4379.

Jörg Waitelonis, Henrik Jürges, and Harald Sack. 2019. Remixing entity linking evaluation datasets for focused benchmarking. *Semantic Web* 10(2):385–412. https://doi.org/10.3233/SW-180334.