NIFify: Towards Better Quality Entity Linking Datasets

Henry Rosales-Méndez
DCC, University of Chile
hrosales@dcc.uchile.cl

Aidan Hogan
IMFD; DCC, University of Chile
ahogan@dcc.uchile.cl

Barbara Poblete
IMFD; DCC, University of Chile
bpoblete@dcc.uchile.cl

ABSTRACT

The Entity Linking (EL) task identifies entity mentions in a text corpus and associates them with a corresponding unambiguous entry in a Knowledge Base. The evaluation of EL systems relies on the comparison of their results against gold standards. A common format used to represent gold standard datasets is the NLP Interchange Format (NIF), which uses RDF as a data model. However, creating gold standard datasets for EL is a time-consuming and error-prone process. In this paper we propose a tool called NIFify to help manually generate, curate, visualize and validate EL annotations; the resulting tool is useful, for example, in the creation of gold standard datasets. NIFify also serves as a benchmark tool that enables the assessment of EL results. Using the validation features of NIFify, we further explore the quality of popular EL gold standards.

CCS CONCEPTS

• **Information systems** → *Information extraction.*

KEYWORDS

Information Extraction; Entity Linking; Benchmark Dataset

ACM Reference Format:
Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2019. NIFify: Towards Better Quality Entity Linking Datasets. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.XXXXXXX

1 INTRODUCTION

Entity Linking (EL) involves annotating entity mentions in a text and associating them with a corresponding unambiguous identifier in a Knowledge Base (KB). EL has gained increasing attention in recent years due mainly to the availability of large KBs on the Web (e.g., Wikipedia, DBpedia, Wikidata, BabelNet) that offer unambiguous identifiers and relevant information for a wide range of entities. For instance, in the sentence **S1** *"Jackson won an award as best-selling artist of the 1980s"* an EL system targeting the DBpedia KB should identify *Jackson* as `dbr:Michael_Jackson`[1]; in this way, we know that the text speaks about a famous musician from the U.S. who is also known as the *King of Pop*. EL thus helps to build a bridge from unstructured information (text) to (semi-)structured data (KBs). Many applications then rely on EL, including semantic search, semantic annotation, text enrichment, entity summarization, relation extraction, and more besides.

Several EL systems have been proposed thus far, along with a range of gold standards for evaluation purposes (surveyed later in Table 1). However, as research on EL has continued to advance, more specialized requirements are being considered, reflecting real environments that stand to benefit from EL; such requirements include multilingualism, specific domains, noisy texts, short texts, semi-structured inputs, etc. With this diversification of requirements, traditional gold standards are not enough: novel gold standards are ideally required to reflect different contexts.

Gold standard datasets are commonly built manually by expert humans reflecting a ground truth. Early datasets were written in (varying) ad hoc formats that required special processing. Hellmann et al [6] thus proposed the NLP Interchange Format (NIF) in order to improve the interoperability of NLP tools, including EL tools. NIF is based on the RDF data model, defining a vocabulary in OWL for representing and sharing NLP-related annotations.

Despite the benefits of NIF, the creation of gold standards is still a complex, error-prone and time-consuming work; hence a number of tools have been proposed to help experts in this task. Röder el al. [17] craft three NIF datasets from texts written in English and German that were tagged manually using their own tool, but to the best of our knowledge the tool is not openly available. Looking for mistakes in datasets, Kunal et al. [9] propose guidelines to validate EL datasets, providing the EAGLET system that checks a variety of quality rules, helping experts to reduce errors; however, some important errors, such as verifying that the target of a link is not a redirect page, are not covered. On the other hand, other works have focused on standardizing the assessment process, providing benchmarking suites (e.g., GERBIL [20], Orbis [15]) that can quickly compare results for state-of-the-art EL systems against a variety of datasets. More generally, all of these NIF operations – creating, validating and performing experiments with EL datasets – have, to the best of our knowledge, been addressed as independent systems.

In this short paper, we thus describe NIFify: a tool that simultaneously supports the creation, visualization, and validation of NIF datasets, as well as the comparison of EL systems. With our tool – shown in Figure 1 – we include some functionalities not covered by previous approaches for creating, modifying and validating NIF datasets. Additionally, we allow to visualise the results of EL systems at both a sentence and document level.

2 BACKGROUND

The typical way to evaluate EL systems is through gold standard datasets, which contain text corpora and their corresponding annotations of entity mentions with respect to the identifiers of a given KB (or multiple KBs). One can then use such datasets in order to measure the quality of the output of an EL system. As more and more such datasets were proposed for EL, interoperability became

[1] Throughout, we use well-known prefixes according to http://prefix.cc

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.XXXXXXX

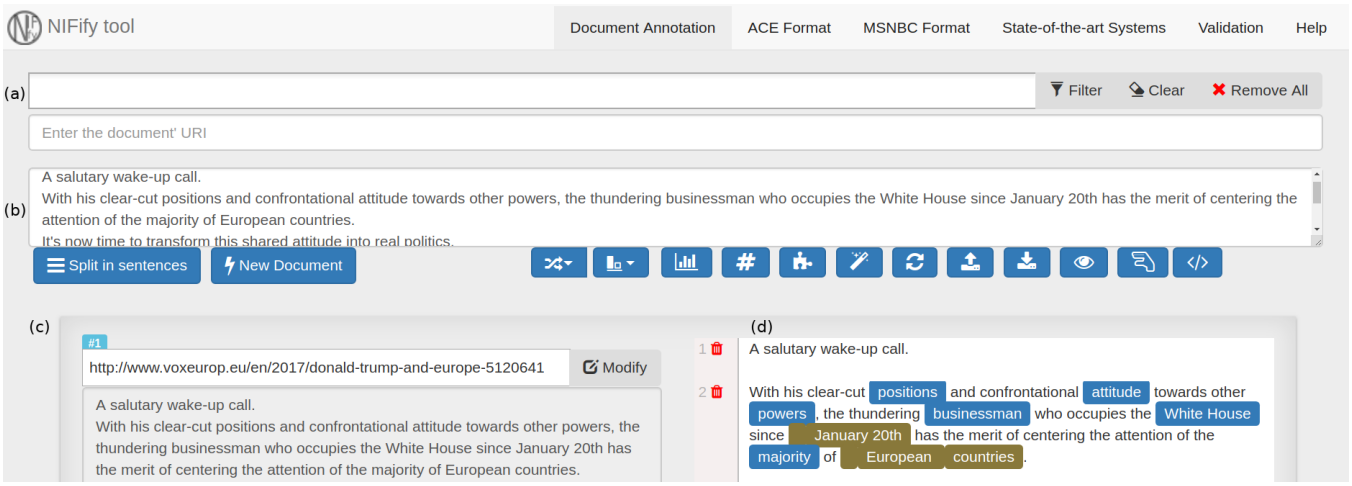

Figure 1: The main view of NIFify showing: (a) the class-reference input to filter annotations; (b) the document text input; (c) the mention identification field; and (d) the annotation visualization.

an issue: various formats were used to represent such datasets. One of the first formats proposed for EL annotation was for the MSNBC [4] dataset, which has two separate files: one a plain text file, and the other an XML file describing the annotations. This same format was followed by other authors proposing further EL datasets. e.g., ACE2004 [16], AQUAINT [16], IITB [10].

However, other EL datasets began to follow other formats. In Table 1 we list some of the most popular EL datasets in the literature along with some details of their content: whether or not they were created manually (**Mn**), whether or not the entity mentions are explicitly typed (**Typ**), and the format used. In terms of formats, many are based on XML (e.g., MSNBC [4], IITB [10], RENDEN [1], CAT [12]) or CSV (e.g., AIDA [7], SemEval [13]). However, a number also use RDF as a base data-model: Melo et al. [5] proposed Lexvo[2] as a RDF-based format and service that defines a unique URI for terms, languages, scripts, and characters from a text corpus; later, Hellmann et al. [6] the NLP Interchange Format (NIF), based on RDF, which is interoperable with a variety of NLP tools, and has been used by several recent EL datasets (e.g., N3-RSS 500 [17], Reuters 128 [17], Wes2015 [22], News-100 [17], DBpedia Abstracts [2], VoxEL [18]). Further legacy datasets were transformed to NIF, including KORE50 and DBpedia Spotlight[3].

NIF is based on RDF triples <*subject*, *predicate*, *object*> where the *subject* identifies a unit of information, such as a document, sentence, or annotation; and each *predicate*—*object* pair defines values for their properties. Figure 2 provides a brief example of a single entity annotation serialized in the Turtle syntax of RDF. The properties `nif:beginIndex` and `nif:endIndex` indicate the start and end position of the entity mention in a sentence; the targeted KB identifier is specified using the property `itsrdf:taIdentRef`; and a class can be defined with `itsrdf:taClassRef`. Other NIF properties capture metadata for other NLP tasks, such as stemming (`nif:stem`), part-of-speech tagging (`nif:oliaCategory`, `nif:lemma`), etc.

[2]http://lexvo.org/ontology; January 1st, 2019.
[3]http://apps.yovisto.com/labs/ner-benchmarks; January 1st, 2019.

Table 1: Overview of popular EL datasets; we highlight in bold those datasets that have been converted to NIF

Dataset	Mn	Typ	Format
MSNBC [4]	✗	✗	MSNBC
IITB [10]	✓	✗	IITB
AIDA/CoNLL [7]	✓	✗	AIDA
ACE2004 [16]	✗	✗	MSNBC
AQUAINT [16]	✗	✗	MSNBC
DBpedia Spotlight [11]	✓	✗	Lexvo
KORE50 [8]	✓	✗	AIDA
N3-RSS 500 [17]	✓	✗	NIF
Reuters 128 [17]	✓	✗	NIF
News-100 [17]	✓	✗	NIF
Wes2015 [22]	✓	✗	NIF
SemEval 2015 Task 13 [13]	✓	✗	SemEval
Thibaudet [1]	✗	✓	RENDEN
Bergson [1]	✗	✓	RENDEN
DBpedia Abstracts [2]	✗	✗	NIF
MEANTIME [12]	✓	✓	CAT
VoxEL [18]	✓	✗	NIF

Figure 2: NIF triples to specify the annotation of Jackson from sentence S1

```
<https://example.org/doc1#char=0,7> a nif:String,
    nif:Context, nif:Phrase, nif:RFC5147String;
    nif:anchorOf """Jackson"""^^xsd:string ;
    nif:beginIndex "0"^^xsd:nonNegativeInteger ;
    nif:endIndex "7"^^xsd:nonNegativeInteger ;
    itsrdf:taIdentRef </wiki/Michael_Jackson> .
```

3 NIF CONSTRUCTION

A number of EL datasets have either been computed from existing sources, or computed automatically. For example, DBpedia Abstracts is too large for human labeling to be feasible.[4] On the other hand, the recently proposed BENGAL tool [14] adopts a creative strategy for automatically generating gold standard datasets: rather than start with text, the authors propose to start with facts about entities from structured datasets (in RDF) and use verbalization components to convert these facts to text, recording which entities are used to generate which sentences; while this approach has the benefit of being able to generate very large and accurate gold standards, how representative the generated text is of real-world corpora depends on the quality of the verbalization component.

On the other hand, per Table 1, most datasets are constructed with manual intervention, and a number of systems have been proposed to help in this process. In previous work, we manually annotated a multilingual EL dataset called VoxEL [18], generating NIF annotations; at the start of this process, we tried to find an existing tool that would aid in the annotation process, but we found that while some systems were unavailable, others (e.g., QRTool[5]) we could not install, or did not offer features such as validation.

Addressing these limitations, we propose NIFify: an open source tool that provides end-to-end support for EL annotation, including the import of text corpora[6]; the import (including the conversion of MSNBC formats to NIF) of existing EL datasets; the addition and revision of annotations; custom tagging systems for annotations; visualizations of annotations; overlapping mentions; and finally, visualisations of the results of EL systems over the resulting dataset. The tool requires no installation and can be used either online or offline in a browser[7]. For space reasons, rather than describe all the features of NIF, we focus on two group of features of particular importance to NIFify: *validation* and *result visualization*.

4 VALIDATION

Validation is a crucial step to help human experts ensure the production of a ground truth for gold standards, and EL datasets are no exception. Legacy EL datasets have been observed to contain errors or design choices that may affect the results of evaluation [9, 19, 21]; furthermore, target KBs may evolve, rendering some links obsolete.

Erp et al. [21], analyze characteristics of seven EL datasets and find biases introduced by the decisions taken in the annotation process; they highlight the need for a more standard creation of datasets. Jha et al [9] propose a set of validation rules and propose the EAGLET system to check these rules when constructing EL datasets; however, these rules are sometimes dogmatic, considering, for example, overlapping mentions to be errors when they are considered valid by other definitions [19]; furthermore, EAGLET requires execution on a command-line to highlight errors in the visualization, rather than being supported by the interface.

NIFify allows for detecting possible errors present in terms of the mentions and the identifiers to which they are linked; specifically, the following rules are checked:

[4]Details of the annotation process are not provided, but we assume it uses links already present in the corresponding Wikipedia texts.
[5]https://github.com/dice-group/QRTool; January 1st, 2019
[6]https://users.dcc.uchile.cl/~hrosales/MSNBC_ACE2004_to_NIF.html; Jan. 1st, 2019
[7]https://github.com/henryrosalesmendez/NIFify_v2; January 1st, 2019

Table 2: Errors found in current NIF datasets; the last dataset was labeled by us

Dataset	SE	LE	FE	CE
DBpedia Spotlight	8	23	4	–
N3-RSS 500	1	34	–	–
Reuters 128	4	71	–	–
News-100	9	1515	–	–
Wes2015	–	609	–	–
VoxEL	–	8	–	–

- Spelling Error (SE): Mentions should neither start nor end in the middle of a word.
- Link Error (LE): When linking to Wikipedia or DBpedia, identifiers should be the URLs/IRIs corresponding to an unambiguous, non-redirect page on Wikipedia.
- Format Error (FE): We check the consistency of the NIF representation with two sub-rules:
  - Annotations are typically assigned a subject IRI of the form `http://example.org#char=`x,y, where x and y should correspond with the values given for `nif:beginIndex` and `nif:endIndex` respectively.
  - The substring identified by these positions should correspond with that denoted by the `nif:anchorOf` property.
- Category Error (CR): For those datasets with classes specified by the predicate `itsrdf:taClassRef`, NIFify allows the specification of custom rules in order to detect inconsistencies in the annotation classes. For example, the classes `dbo:Person` and `dbo:Event` should not be present on the same annotation as they are disjoint: an entity is typically not a person and an event at the same time.

NIFify then encodes rules to detect these errors and thus validate EL datasets. In order to test the prevalence of these errors in existing datasets, we ran NIFify's validation over EL datasets currently available in the NIF format (excluding those that we converted ourselves to NIF – MSNBC and ACE2004 – since we resolve such errors as part of the conversion). In Table 2, we show the results of this validation process, where we can observe that all datasets considered contain errors of at least one type.

In the majority of the cases, SE errors are introduced in the construction of the dataset with the addition of characters that do not belong to the mention, or on the contrary, leaving out part of a word that completes a mention; for example, in the DBpedia Spotlight dataset, the URI `wiki:Man` is associated with the three characters of the world *performance*. Other SE errors contained in the datasets involve missing spaces between words.

The most frequent type of error encountered in the NIF dataset was LE: this is mainly due to the fact that KBs are constantly evolving, which may affect link consistency. For example, in Wikipedia, pages about specific entities may become disambiguation pages, or redirects to other pages. Such changes explain why our own dataset (VoxEL, created using NIFify) contains such errors: the external KB has evolved since its creation. The News-100 and Wes2015 contain a large number of LE errors beyond what can be explained by the KB changing: for example, in the Wes2015 dataset, 520 of its LE

errors correspond to redirect pages, 48 to disambiguation pages, while the rest do not point to valid pages.

Finally, the only dataset we found with FE-type errors was DBpedia Spotlight, which had problems with its NIF representation. On the other hand, we did not find any errors of type CE.

We have published all errors found online for reference.[8] We conclude that most of the validation features of NIFify can help to improve the quality of EL datasets, including to find problems caused by the evolution of a KB over time.

5 RESULT VISUALIZATION

Once an EL dataset has been generated, the next step is to evaluate and compare EL systems using the dataset. A number of systems have been proposed to help evaluate and compare EL systems. Cornolti et al. [3] proposed the BAT framework, which they used to compare five EL systems over five datasets. Along similar lines, Usbeck et al. proposed GERBIL [20], which extends the systems and (NIF) datasets supported. However, both frameworks produce comparative metrics, rather than visualizing the actual output of the EL tool(s). Another EL benchmark framework called Orbis [15] was recently proposed that includes visualization of systems' responses; however, Orbis is not available in the provided URL.[9].

Given that there is no clear definition on what EL systems should link [19], we argue that metrics like precision and recall may not tell the full story, and that results may be due not only to the quality of the output produced by an EL system, but also whether or not it targets the same types of entities as labeled in the dataset. Comparing EL results with the ground truth labeled in a dataset under construction/revision may even lead to changes in the dataset.[10] Hence with NIFify we propose a benchmark framework to visualize the results of EL systems over the NIF dataset, highlighting both *true positives* or *false positives*, which allows a more qualitative assessment of both a given EL tool and an EL dataset, possibly in the context of a given application. Additionally, NIFify can be used to demo EL systems, offering a visual, friendly user interface.

6 CONCLUSION

In this short paper, we describe the NIFify system, which aims to address a number of shortcomings of existing tools for generating EL datasets and evaluating EL tools: in particular, NIFify simultaneously supports the creation, visualization, and validation of NIF datasets, as well as the comparison of EL systems. We first discussed some extensions to the NIF format to support mentions having multiple possible identifiers annotated with different types. We then provided a summary of the main features of NIFify for generating EL gold standard datasets, before focusing on features relating to validation, showing that existing EL datasets exhibit errors detectable by the tool, detecting a total of 2,321 errors across six datasets; we publish these errors online for reference: https://users.dcc.uchile.cl/~hrosales/dataset_errors.html. Finally, we discuss the importance of features for visualizing the results produced by an EL system, which are further implemented in the NIFify tool. A demo of the tool is available at https://users.dcc.uchile.cl/~hrosales/NIFify_v2.html

[8] https://users.dcc.uchile.cl/~hrosales/dataset_errors.html; January 1st, 2019.
[9] https://github.com/htwchur; January 1st, 2019.
[10] Of course, we urge caution to ensure that bias is not introduced by adapting a dataset to suit a subset of tools evaluated.

7 ACKNOWLEDGMENTS

The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also supported by the Millennium Institute for Foundational Research on Data (IMFD) and by Fondecyt Grant No. 1181896.

REFERENCES

[1] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: named entity linking in digital literary editions using linked data sets. *CSIMQ* 7 (2016), 60–80.
[2] Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. 2016. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus.. In *LREC*.
[3] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *World Wide Web Conference*.
[4] Silviu Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *EMNLP-CoNLL* (2007), 708.
[5] G. de Melo and G. Weikum. 2008. Language as a foundation of the Semantic Web. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference*.
[6] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP Using Linked Data. In *ISWC*. 98–113.
[7] Johannes Hoffart and et al. 2011. Robust disambiguation of named entities in text. In *EMNLP*. ACL, 782–792.
[8] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *CIKM*. 545–554.
[9] 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 *ESWC*. 305–320.
[10] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *SIGKDD*. 457–466.
[11] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *I-SEMANTICS*. ACM, 1–8.
[12] A.L. Minard and et al. 2016. MEANTIME, the NewsReader multilingual event and time corpus. (2016).
[13] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking.. In *SemEval@ NAACL-HLT*. 288–297.
[14] 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 *Proceedings of the 11th International Conference on Natural Language Generation*. 339–349.
[15] Fabian Odoni, Philipp Kuntschik, Adrian M. P. Brasoveanu, and Albert Weichselbraun. 2018. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In *SEMANTICS*. 33–42.
[16] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *NAACL-HLT*. 1375–1384.
[17] 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 *LREC*. 3529–3533.
[18] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2018. VoxEL: A Benchmark Dataset for Multilingual Entity Linking. In *ISWC*. 170–186.
[19] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. 2018. What Should Entity Linking link?. In *AMW*.
[20] Ricardo Usbeck and et al. 2015. GERBIL: General Entity Annotator Benchmarking Framework. In *WWW*. 1133–1143.
[21] Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC*.
[22] Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Linked data enabled generalized vector space model to improve document retrieval. In *NLP & DBpedia @ ISWC*.