Improving Named Entity Linking Corpora Quality

Albert Weichselbraun*, Adrian M.P. Braşoveanu**, Philipp Kuntschik *, Lyndon J.B. Nixon**

* HTW Chur, Switzerland ** MODUL Technology GmbH, Vienna, Austria.

Why Focus on NEL Corpora Quality?

Because evaluations still need high-quality corpora!

- Deep Learning (DL) and Big Data go hand in hand.
- Links and NILs are unstable due to Knowledge Base (KB) evolution.
- KB translation is possible, but corpora are rarely updated!
- Multiple annotation sets can be merged or used to compute different evaluation scores (e.g., weak or strong matches).
- DL requires fast and automated annotators, therefore we need some kind of warranty that they will perform well.

Corpora Publishing Methodology

Standardized structure

- **Corpus folder** containing all data and annotations in multiple formats (e.g., csv, NIF).
- **metadata.yaml** files that describes the current corpus version metadata.
- **README.md** that provides additional general information.
- **Data statements** to describe the intended usage for NLP experiments.
- **Annotation guideline** that describes the rules used by the human or machine annotators during the annotation process.
- **Code** used to generate the data set (if possible or if needed).
- **Revisions history** in order to track the big changes.
- **List of previous versions** in order to enable reproducibility of old papers.

Table 1: Suggested corpus metadata

Metadata	Description
corpus_name	A name that identifies the corpus.
corpus_url	The corpus archive URL.
creator	Comma-separated list of creators.
date	The corpus's publishing date.
description	Description of current corpus version.
final	Is it usable for official evaluations?
parent_corpus_url	The URL of the parent corpus.
considers_corpus_url	List of related corpus versions.
annotation_style	A list of annotation styles per supported entity types.
annotators_per_document	Number of annotators per document.
annotator_agreement	Inter-rater-agreement between annotators.

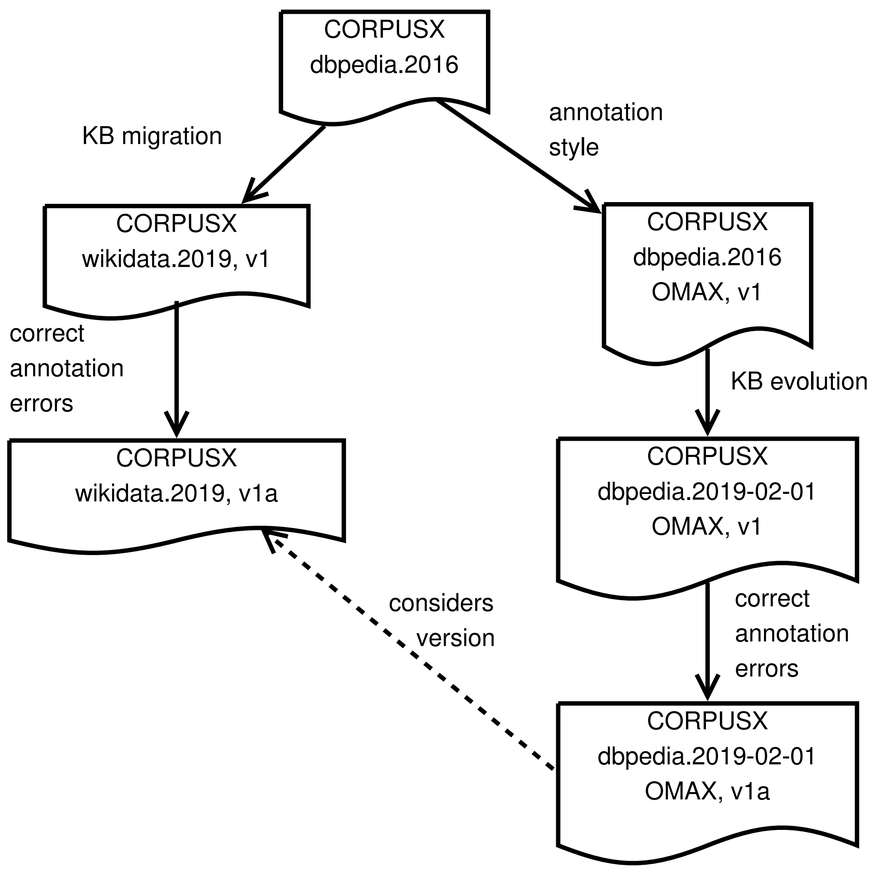

Figure 1: Common corpus versioning use cases.

Frequent NEL Evaluation Errors

- **Data set (DS) errors** are those produced during the annotation process.
- **Knowledge Base (KB) errors** are generally caused by wrong attributes or KB evolution.
- **Annotator (AN) errors** are caused by the evaluated system.
- **Scorer Errors (SE)** are caused by the evaluation tool.

Gold

Iran decided to continue its operations east of Basra to destroy Iraqs forces, Iranian Prime Minister Mir-Hossein Mousavi told Tehran Radio. That seems to be the best spot for the complete destruction of (Iraqi President) Saddams forces. This is why it was decided that the operations would forcefully continue there, he said. At the same time, we have maintained our ability to act throughout the length of the fronts, Mousavi told the radio, monitored by the British Broadcasting Corporation. Tehran Radio reported over 1,500 Iraqi casualties today as Iran continued its Karbala-8 operation launched early yesterday. Iran said over 2,600 Iraqis had been killed or wounded yesterday. Iraq said today its forces had beaten back the latest attacks inflicting heavy casualties. Mousavi told Tehran radio that on the whole, our advances have broken the back of the military forces of Saddam. Earlier today the Iranian news agency IRNA said Tehrans forces were stabilising new positions after their assault on Iraqi lines defending Basra, Iraqs second city. In Baghdad, a military spokesman said Iraqi warplanes today destroyed an oil pumping station and a production unit at Irans Ahvaz field.

Predicted

Iran decided to continue its operations east of Basra to destroy Iraqs forces, Iranian Prime Minister Mir-Hossein Mousavi told Tehran Radio. That seems to be the best spot for the complete destruction of (Iraqi President) Saddams forces. This is why it was decided that the operations would forcefully continue there, he said. At the same time, we have maintained our ability to act throughout the length of the fronts, Mousavi told the radio, monitored by the British Broadcasting Corporation. Tehran Radio reported over 1,500 Iraqi casualties today as Iran continued its Karbala-8 operation launched early yesterday. Iran said over 2,600 Iraqis had been killed or wounded yesterday. Iraq said today its forces had beaten back the latest attacks inflicting heavy casualties. Mousavi told Tehran radio that on the whole, our advances have broken the back of the military forces of Saddam. Earlier today the Iranian news agency IRNA said Tehrans forces were stabilising new positions after their assault on Iraqi lines defending Basra, Iraqs second city. In Baghdad, a military spokesman said Iraqi warplanes today destroyed an oil pumping station and a production unit at Irans Ahvaz field.

Gold Entities

Iran (http://dbpedia.org/resource/Iran): 0 - 4: Place

Basra (http://dbpedia.org/resource/Basra): 48 - 53: Place

Mir-Hossein Mousavi (http://dbpedia.org/resource/Mir-Hossein_Mousavi): 102 - 121: Person

Mousavi (http://dbpedia.org/resource/Mir-Hossein_Mousavi): 420 - 427: Person

British Broadcasting Corporation (http://dbpedia.org/resource/BBC): 461 - 493: Organization

Iran (http://dbpedia.org/resource/Iran): 554 - 558: Place

Iran (http://dbpedia.org/resource/Iran): 619 - 623: Place

Iraq (http://dbpedia.org/resource/Iraq): 685 - 689: Place

Mousavi (http://dbpedia.org/resource/Mir-Hossein_Mousavi): 776 - 783: Person

Saddam (http://dbpedia.org/resource/Saddam_Hussein): 881 - 887: Person

Predicted Entities

Iran (http://dbpedia.org/resource/Iran): 0 - 4: Place

Basra (http://dbpedia.org/resource/Basra): 48 - 53: Place

forces (http://dbpedia.org/resource/International_Security_Assistance_Force): 71 - 77: Organization

Mir-Hossein Mousavi (http://dbpedia.org/resource/Mir-Hossein_Mousavi): 102 - 121: Person

Tehran (http://dbpedia.org/resource/Tehran): 127 - 133: Place

forces (http://dbpedia.org/resource/International_Security_Assistance_Force): 230 - 236: Organization

Mousavi (http://dbpedia.org/resource/Mir-Hossein_Mousavi): 420 - 427: Person

British Broadcasting Corporation (http://dbpedia.org/resource/BBC): 461 - 493: Organization

Tehran (http://dbpedia.org/resource/Tehran): 495 - 501: Place

Iraqi (http://dbpedia.org/resource/Iraq): 528 - 533: Place

Figure 2: Gold versus annotator without NILs for Reuters-128's document 107 displayed in Orbis

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

Figure 3: Common errors in NEL corpora

Lenses: An Alternative for Improving Quality

Table 2: 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 3: 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	$m^{e_i,KB'}_{[x_i,y_i]}$	$m^{nil,KB'}_{[x_i,y_i]}$	$m^{e_{i1},KB'}_{[x_{i1},y_{i1}]}, \ldots, m^{e_{in},KB'}_{[x_{in},y_{in}]}$	$m^{e_i,KB'}_{[x_i,y_i]}$

Table 4: 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	$m^{\emptyset}_{[s'_i]}$	$m^{\emptyset}_{[s'_i]}$

- **No single correct way of annotating a document. What should we do?**. Multiple annotation sets can sometimes provide a solution!

Acknowledgements

ReTV EPOCH MedMon MODUL TECHNOLOGY HTW Chur Hochschule für Technik und Wirtschaft University of Applied Sciences