Framing Named Entity Linking Error Types

Adrian M.P. Braşoveanu*, Giuseppe Rizzo‡, Philipp Kuntschik†, Albert Weichselbraun†, Lyndon J.B. Nixon*
* MODUL Technology GmbH,
Am Kahlenberg 1, 1190 Vienna, Austria
‡ ISMB, Via Pier Carlo Boggio 61, 10138 Torino, Italy
† Swiss Institute for Information Research, University of Applied Sciences Chur,
Pulvermühlestrasse 57, 7000 Chur, Switzerland
{adrian.brasoveanu,lyndon.nixon}@modul.ac.at
giuseppe.rizzo@ismb.it
{philipp.kuntschik,albert.weichselbraun}@htwchur.ch

Abstract

Named Entity Linking (NEL) and relation extraction forms the backbone of Knowledge Base Population tasks. The recent rise of large open source Knowledge Bases and the continuous focus on improving NEL performance has led to the creation of automated benchmark solutions during the last decade. The benchmarking of NEL systems offers a valuable approach to understand a NEL system's performance quantitatively. However, an in-depth qualitative analysis that helps improving NEL methods by identifying error causes usually requires a more thorough error analysis. This paper proposes a taxonomy to frame common errors and applies this taxonomy in a survey study to assess the performance of four well-known Named Entity Linking systems on three recent gold standards.
Keywords: Named Entity Linking, Linked Data Quality, Corpora, Evaluation, Error Analysis

1. Introduction

A Named Entity Linking (NEL) system identifies, classifies and links entity mentions from a text to their Knowledge Base (KB) references. A NEL system can also be used for extracting factual knowledge from a text in order to use it for Knowledge Base Population (KBP). The typical components of a NEL system reflect the logical steps of identifying and linking the entities: identification and classification (to a defined type like person, organization or location) of the entity mentions in a text (Named Entity Recognition and Classification or NERC), linking to the referent KB, and clustering of the remaining unlinked entities (Ji and Nothman, 2016). Some of these steps might be performed jointly if the chosen architecture supports it (e.g., neural architectures). Numerous current architectures used for NEL systems include: *(i)* graph-based disambiguation which uses the links between the entities found in the text in order to rank the best candidates; *(ii)* statistical disambiguation typically focused on classic machine learning algorithms or heuristics (e.g., SVM, Conditional Random Fields, etc.) or *(iii)* neural models (e.g., LSTM). When evaluating a system, at least three other components are needed: a dataset (usually a labeled corpus or gold standard), a certain KB version (e.g., DBpedia 2016-04), and a scorer that computes measures such as accuracy, precision, recall, and F-measure.

These measures compare the NEL systems quantitatively, but do not help system designers improve their approaches. If, for example, one wants to improve the precision of a system because it affects its perceived quality, the number of false positives needs to be decreased. Reducing the number of errors (false positives and false negatives) requires a thorough analysis of the evaluation results. A good method to do this is to use the primary error analysis results from the TAC-KBP inspired neleval[1] scorer (Hachey et al., 2014), which is limited to indicating whether the returned results are correct, incorrect, extra (i.e. a named entity does not occur in the gold standard) or missing. If possible, a more detailed explanation of each error should be included, as it could potentially lead to the rapid updating of systems, datasets and KBs, especially since all these components could potentially trigger errors during an evaluation. In fact, since some of these components will not necessarily be under the system developer's control as gold standards, KBs, or scorers are probably developed by third-parties, we might argue that such an explanation is not only beneficial, but that it should become standard practice.

Given the increasing complexity of the NEL systems and evaluations, as a first step towards an automated error classification system, we propose a taxonomy focused around a set of types (e.g., Knowledge Base, Dataset, Annotator, NIL Clustering, Scorer, etc.) and sub-types (causes) as reflected through the various steps of a NEL system (e.g., partial match, wrong link returned due to KB redirect). Such a classification effort allows us to perform multiple tasks when processing the results of the error analysis: *(i)* create a transparent and reproducible method to publish detailed evaluation results[2] on top of the well-known TAC-KBP scorer; *(ii)* evaluate and improve the quality of the labelled data (e.g., Knowledge Base dump, gold standards) used in the evaluation; *(iii)* help NEL system designers to improve their systems by fixing the errors at the level where they are produced (e.g., we can report the Knowledge Base or gold standard errors to the creators of these resources).

The remainder of this paper is organized as follows: Section 2. discusses related work; Section 3. presents the fram-

[1] https://github.com/wikilinks/neleval

[2] The annotation guideline can be found at https://github.com/modultechnology/nel_errors/tree/master/guideline.

ing of the problem and the reasoning process behind the proposed taxonomy; Section 4. applies this taxonomy in a survey study aimed at analyzing different gold standards and NEL systems using DBpedia as the referent KB; and Section 5. concludes the paper and outlines future research.

2. Related Work

TAC-KBP challenges (Ji and Nothman, 2016) are focused on the rapid prototyping of NEL systems for various languages. Each participant needs to submit one or several runs of their system and write a paper. In addition to the new yearly datasets, TAC-KBP participants are given access to previous datasets for training purposes. The event's annual overview reviews the best approaches and shortlists the challenges to be solved for the next years. The goal is to find approaches to solve these errors in next year's competition. In the 2016 edition's overview (Ji and Nothman, 2016), ample space was given to the approaches used for trilingual knowledge transfer, weak/strong mention boundary detection and to the within-document and cross-document coreference resolution error propagation, whereas type discovery (e.g., discovering entities that are not in the current schema), massive multilingual Entity Discovery and Linking (e.g., for hundreds of languages instead of just three) or streaming data were considered as challenges for 2017. It has to be noted that one of the most popular scorers used today (neleval) is based on the lessons learned from the TAC-KBP challenges (Hachey et al., 2014). Radford (Radford, 2015), a former TAC-KBP participant, analyzed his systems and presented types of errors encountered in them, but tailored its taxonomy of errors around the peculiarities of each system.

Issues observed in well-known gold standards are reported in (van Erp et al., 2016), but the results are described through a general set of features (such as dataset types, confusability, prominence) and not through the results obtained by different NEL systems. The paper also notes that the efficacy of NEL evaluations needs to be improved by removing the dependency on black-box evaluations (e.g., evaluations in which the participants only see global results and not mention-level results) as they do not allow to improve upon the results. We address this research challenge by building our experiments on top of the TAC-KBP scripts and evaluation format (Hachey et al., 2014) to provide a more detailed error analysis.

In (Heinzerling and Strube, 2015), authors present a system for performing visual error analysis built on top of the TAC-KBP evaluation formats, but it only supports a subset of the error types covered by our method.

In (Cornolti et al., 2013), authors define a set of annotation tasks (such as Annotate to Wikipedia – A2W, Concepts to Wikipedia – C2W) and propose an automated evaluation system that measures per-task performance. A sequel to Cornolti's work, GERBIL (Usbeck et al., 2015) is a large-scale evaluation system that allows us to compare the output of different NEL systems. It is typically used as an alternative to TAC-KBP style evaluations in literature. GERBIL uses gold standards in the NIF format (Natural Language Processing Interchange Format), an RDF format designed to allow the sharing of both textual and annotation resources and ease the interplay between NLP tools. A basic error analysis based on Gerbil can be performed with EAGLET, but it is only focused on seven error types which can be classified as dataset errors (Jha et al., 2017) and does not include other large error classes. The EAGLET pipeline contains a preprocessing module and a rule-based module for identifying dataset errors, resulting annotations being reviewed by a human annotator. It has to be noted that similarly to our approach, the observed errors are annotated and judged by humans. Two of the gold standards we experiment in our work (KORE50 and Reuters128) are also integrated in GERBIL[3].

3. Classification of Errors in NEL Evaluations

Current NEL scorers, such as GERBIL or TAC-KBP do not allow to perform a closer inspection and framing of the evaluated NEL annotations. While GERBIL does not provide any mechanism through which to access the errors at mention level, the TAC-KBP results can be processed to obtain a primary error analysis limited to the validity of the results (e.g., results are marked as wrong link, extra link or missing). While such information is valuable, adding a semantic layer on top of it can offer researchers and developers the insights they need to improve their tools. In order to capture the logical reasoning that has produced the error we have considered the following error types:

KB A Knowledge Base error is an error discovered in a particular KB version (e.g., DBpedia 2015-10). May include wrong mappings (e.g., a person's name that points to a year) or missing entities.

DS Dataset errors are typically the errors produced by the human annotators during the annotation process which can still be found in a certain gold standard version. Most common DS errors are missing or wrong annotations (e.g., incomplete surface form).

AN Annotator errors refer to the output of an automated annotation system that follows the classic NEL phases (e.g., NERC, linking, relation extraction or graph disambiguation). This is the largest error class and includes errors like wrong type, wrong link, wrong surface form, etc.

NIL NIL Clustering errors refer to the output of the NIL Clustering components. Known errors include missing surface forms for some of the entities or new links in recent version of the KB.

SE Scorer or evaluation errors explain the errors reported by an evaluation script when the AN and DS outputs are similar. These types of errors are extremely hard to spot. A common error is related to how redirect links are scored.

We have started our investigations by examining the output of the TAC-KBP scorer. We have followed the strategy of collecting and verifying the mentions (surface forms), types

[3] http://gerbil.aksw.org/gerbil/config

NEL System			Gold Standard		Error	
Entity Link$_s$	**ET$_s$**	**SurfaceForm**	**Entity Link$_g$**	**ET$_g$**	**Type**	**Cause**
dbr:Bruce_Willis	ORG	expiration	-	-	KB	Redirects
de.dbr:2009	LOC	2009	-	-	KB	Wrong Type
dbr:United_States	LOC	U.S.	-	-	DS	Missing Annotation
dbr:New_York_City	LOC	New York	dbr:New_York	LOC	DS	Wrong Annotation
de.dbr:Berlin	LOC	Berlin	dbr:Berlin	LOC	DS	Different Language
dbr:JFK	PER	Kennedy	dbr:JPK	PER	AN	Same-Type
dbr:Beck	ORG	Beck	dbr:Jeff_Beck	PER	AN	Cross-Type
dbr:Barack_Obama	PER	Malia Obama	NIL	PER	NIL	Wrong Cluster
NIL	ORG	Knicks	dbr:New_York_Knicks	ORG	NIL	Partial Match
dbr:Miles_Davis	PER	Davis	dbr:Miles_davis	PER	SE	Correct Redirect

Table 1: Examples of the most common errors detected in the three gold standards investigated in this paper. The entries that have gold links were marked as *wrong-link* and the others as *extra* in TAC-KBP primary error analysis. ET represents the entity type. Subscripts *s* and *g* denote the *system* (annotator) and *gold standard* (dataset).

and links for all entities present in a text. The template used for describing a new error cause in the *Annotation Guideline* contains the **scope** (mention, type, link), **similarity** to other error causes, a **general description** of the error, **examples** and **comments**. We aim to improve this template in time using community feedback. Our initial goal was simply to find an easy way to report such errors through a method that would later allow us to easily classify them.

Based on the experiments performed in Section 4., we have defined error causes for each error class as illustrated in Table 1. In a first phase we have focused on the first two large classes: KB and DS in order to remove any doubts related to the AN or NIL errors.

A typical **KB** error looks like the entity *de.dbr:2009* that has been marked as a location in the German DBpedia version 2015-10 (Table 1) or the surface form *expiration* that was tagged with *dbr:Bruce_Willis* due to a KB redirect. A lot of the errors that occur on this level are simply due to the fact that most systems do not use the live versions of the KB but rather dumps that are published at certain intervals (e.g., DBpedia dumps are published every 6 months, whereas Wikidata dumps are published weekly). It is quite often the case that missing links from a previous dump were updated in the meantime.

The **DS** errors are generally instances of wrong annotations due to various causes: typos (we found many cases in which dots were missing from geographic abbreviations), a different language than the target one (e.g., German DBpedia instead of English), partial matches (e.g., geo entities missing parts of their name). DS errors are perhaps the hardest to agree on as each gold standard could have different annotation guidelines. However, we think these must be judged both against the original guideline that was used to create them, but also using common sense, especially if there is an intention of integrating multiple datasets in a single evaluation tool (e.g., as it was done with GERBIL).

AN errors include abbreviation conflicts (e.g., for *Kent.* abbreviation, the annotator returns *Kent, UK* instead of *Kentucky*), same-type disambiguation errors (e.g., *Bill Clinton* returned instead of *Bill Gates* or *Hillary Clinton*), cross-type disambiguation errors (e.g., when an entity with a different type is returned), or generic terms (e.g., when words like *ship* or *Admiral* are returned instead of the real entities that are near them, like *Hansa Stavanger container ship* or *Admiral Thad W. Allen*). Largely the AN errors depend on the algorithms and settings that were chosen for a specific tool. These kind of errors can only be removed by fixing the tool.

NIL clustering errors include entity mentions being shared among multiple clusters (e.g., role/title or last name appearing in a different cluster than the full name of a person) or partial macthes (e.g., *Knicks* used for *N.Y. Knicks*). Or clusters that are generated of lexical equal entities but they hold different semantic meanings in the context they are used. This large error class is, unfortunately, dependent on the annotation system. Early systems like those presented in Radford's work (Radford et al., 2011) were known to be sub-optimal due to their lack of integration with the linking process, whereas more recent implementations are integrative and also include advanced co-reference resolution algorithms as described in recent TAC-KBP initiatives (Ji and Nothman, 2016).

While **SE** errors are not as frequent like the other categories, a classic example is represented by correct redirects not counted as such (see the Miles Davis row in Table 1). In order to correctly identify such errors there is a need to compare the results of different evaluation tools on the same datasets, but this goes beyond the purpose of the current paper.

It has to be noted that in some cases an error can appear due to multiple causes (e.g., a partial match causes a different entity to be returned by the NEL system due to a wrong KB redirect - this case offering both an AN and a KB error). Such cases are of course hard to interpret correctly, but we have chosen to follow the logical order: KB - DS - AN - NIL - SE and try to always place the error on the first layer on which it can occur. This might not always be optimal, but it should help developers better reason about how these errors are produced.

The reasoning process used to identify such cases starts with reading the text in order to understand the context. Then the surface forms, types and mentions that are present

in AN and DS results are examined in order to find the most likely error cause (e.g., surface form does not seem to have any connection with the returned entity, wrong link or wrong type). If the error cause cannot be found in AN or DS results, then the KB entry is examined for additional clues. This reasoning chain is repeated until a good error cause can be determined, and if the error cause is not present in the Annotation Guideline, a new entry is added to this guideline and is formatted according to the proposed template.

4. Experimental Setup

A set of experiments was performed on three gold standards with four NEL systems in order to better understand the reasoning used for explaining the errors, the agreement among human annotators who evaluated the system results, and the feasibility of creating an automatic semantic error analysis system for NEL evaluations.

4.1. Datasets and Tools

We have only selected datasets that were known to have been annotated manually and were published in the NIF format. If the dataset publication mentioned that it was created automatically, it had only one annotator or is rather a baseline than a gold standard, we have not included it in our experiment. While it can be argued that the methodology described in the previous section can also be applied for such datasets, we thought it is best to first select datasets that match our criteria for reasonable gold standards. We consider extending this methodology for any type of datasets in future work.

We have applied the described methodology for identifying and classifying the error classes and error types introduced in the previous section to evaluations performed on four state-of-the-art off-the-shelf NEL systems, namely DBpedia Spotlight (Daiber et al., 2013), Babelfy (Moro et al., 2014), AIDA (Hoffart et al., 2011), and Recognyze (Weichselbraun et al., 2015) while annotating three well-known benchmark datasets, namely Reuters128 (Röder et al., 2014), KORE50 (Hoffart et al., 2012), and RBB150 (Brasoveanu et al., 2016).

AIDA and Babelfy are graph-disambiguation frameworks, DBpedia Spotlight is a statistical disambiguation framework, whereas Recognyze used heuristics at the time of the experiments. We have selected two graph-disambiguation frameworks simply because it seemed to be the best paradigm for performing NEL at the time.

KORE50 consists of English sentences from five different domains, Reuters128 includes full news media articles in English, and RBB150 contains German television subtitles. To reduce the manual annotation workload, we processed a subset of documents for each corpus and focused only on the false positives.

4.2. Experimental Methodology

We have then used the four NEL systems in order to automatically annotate the first 50 texts from each gold standard and collect the false positives that could signal eventual problems. DBpedia Spotlight was the only annotator used for both languages, while Recognyze was only used for German. The gold standards were converted into the TAC-KBP format from the NIF format (Hellmann et al., 2012), a format that allows for easy interchange of NLP data (e.g., gold standard annotations, NEL system results). The TAC-KBP scorer (Hachey et al., 2014) was used for the evaluations. A set of runs with the webservices of the investigated NEL systems was produced for each gold standard. Only three types presented in TAC-KBP 2014 evaluations were selected: Person, Organisation and Location as we considered that the systems were already well-trained to handle them. TAC-KBP 2015 has introduced the convention of splitting the Location class into separate classes for GPE, Location and Facility, but all the other challenges still consider only few types. The types of the entities were inferred from the DBpedia links returned by the systems, as not all systems return the types directly. We have chosen to keep all well-known classes for the three entity types as they appear in DBpedia, YAGO and schema.org ontologies and have not included fine-grained typing in the current version. While fine-grained typing is in our research agenda, our current goal is to refine this methodology based on feedback received from third-party users.

We checked all the entities that were marked as having no other type than *owl:Thing* and discovered that in some cases they represented the merging of multiple entities (e.g., *Kenneth and Mamie Clark*), family names (e.g., *dbr:Reuter*), redirects (e.g., *Carnival Cruise Lines*) or even annotations in other languages (e.g., links from German DBpedia in an English corpora). Only entities that had the three types we were aiming for (person, location, organization) or were annotated to different languages in the gold standard were considered. We have then ran the TAC-KBP scorer and collected the results and the false positives from the primary error analysis (the classification of links as correct, wrong, missing or extra returned by neleval) for each run.

We automatically created supersets with all the identified errors available for a particular gold standard. Such a superset included all the data related to the errors available from the NEL systems and gold standard, each error being identified by the mention, type and link from both the AN and DS (if available), but also by several fields that we have later used to create our error annotations (e.g., fields like document, span, error type, error cause or presence in the gold standard). These supersets were annotated independently by two human annotators and inter-annotator agreement scores were computed. A third human annotator went through the results along with the two human annotators and resolved the inconsistencies. While creating human annotations is costly, we considered it a necessary step in order to validate our taxonomy, but also to create a gold standard that can be used towards the automated classification of errors.

All the errors were annotated with respect to single-language evaluations (e.g., English, German). While multilingual evaluations start to become more common, there are less datasets for such tasks available, and most of them are rather baselines then gold standards.

Along with this paper we also publish the human annotation guidelines to foster their reuse, to promote common annotation standards and to advance the state of the art in

Dataset	Spotlight	Babelfy	AIDA	Recognyze
Reuters128	125	172	102	-
KORE50	23	18	44	-
RBB150	113	-	-	65

Table 2: False Positive counts returned by the investigated NEL systems.

Dataset	κ	FPs	KB	DS	AN	SE
Reuters128	0.653	302	9	42	251	0
KORE50	0.689	59	2	1	55	1
RBB150	0.877	176	2	70	104	0

Table 3: Total count of False Positives (FPs) and error types (KB, DS, AN, SE) in all systems.

NEL error analysis[4].

4.3. Results and Discussion

The examples presented in Table 1 were collected during the experiments. Table 2 shows a quantitative analysis of false positives generated by the NEL system. Table 3 reports on the inter-annotator agreement Fleiss's κ (Fleiss, 1971). The Fleiss's κ agreement figures show that there is a high agreement between the human annotators (a value between 0.61 and 0.80 denotes substantial agreement, while a value between 0.81 and 1.00 denotes an almost perfect agreement). NIL Clustering errors were not taken into account in this experiment due to the fact that the examined systems included no such components. As it can be seen only one gold standard can be considered as having high-quality annotations (KORE50), due to its low number of KB errors and high agreement with annotations provided by state-of-the-art NEL tools. The Reuters 128 dataset that is filled with many popular entities, in contrast, suffers from numerous redirects and multiple surface forms that lead to the largest number of KB errors.

The RBB150 gold standard contains the largest number of errors overall, although many of these errors are caused by partial matches due to the fact that annotations for person-type entities included their roles (e.g., President Barack Obama), whereas most NEL systems returned these entities without roles (e.g., Barack Obama), therefore the dataset's annotation rules should be changed to be more in line with other datasets or provide different settings for full and partial matches.

Both Reuters128 and RBB150 have cases of entities annotated in a different language than the original language of the processed text (e.g., German annotations in an English text). This should be considered an error only in a single language evaluation, whereas in a multilingual evaluation such output is not only desirable, but it is encouraged.

We have noticed that entities that are classified as Person lead to a substantial agreement among the NEL systems, while entities that are classified as Organization are often annotated with their full suffixes in the gold standards, but recognized without them. Location proved to be the source of many inconsistencies in gold standards and KB, due to demonyms automatically annotated to countries (therefore Annotator errors) or lack of clear rules for annotating long names (e.g., Columbus, OH might be a single entity, while for Rome, Italy there will be two annotations).

A good application that can help improve the quality of semantic data is the rapid publishing of the evaluation results. Before turning the output into NIF and reporting issues to KB or DS designers, we advise practitioners to establish several rules upfront, for example to clarify how many tools and human annotators would need to be in agreement. While we recommend at least an agreement between 75% of the tools and two human annotators, the final criteria will always depend on the use case and end goals of system or challenge designers.

5. Conclusion

In this paper we described a taxonomy to identify errors in gold standards and errors generated by NEL systems. The taxonomy has been tested in an experimental environment that has involved two human annotators. A manual annotation and classification of the errors identified in the evaluation results demonstrates the usefulness and potential of the suggested schema for identifying error classes and improving the underlying datasets, KBs and NEL components.

While the taxonomy and the evaluation method presented here are still in an early stage (e.g., only false positives were considered and the focus was mostly on the big error classes), a few applications already show lots of promise in using such an inductive data-driven approach of fixing gold standards, knowledge bases, NEL system annotations, and overall providing a better support to error analyses for NEL evaluations.

Using human annotators in order to annotate the errors is a current limitation of the method presented in this paper due to time consumption and costs, but it was necessary as a first step towards automating error analysis. The agreement scores denote the fact that the method can be widely deployed and by converting these annotations in the NIF format we can share our results with the maintainers of KBs and gold standards in order to help them to improve their services.

The main advantage of this method is the fact that it allows us to identify and explain most of the large error classes as long as the NEL systems considered include the conventional components (even if some of these components are merged). This work can also be adapted for different NEL evaluation systems besides TAC-KBP (e.g., GERBIL) with the condition to be able to access the NEL system runs and the gold standards.

6. Acknowledgements

The InVID project (http://www.invid-project.eu/) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687786. The Job-Cockpit project (www.htwchur.ch/job-cockpit) was funded by the Swiss Commission for Technology and Innovation (CTI).

[4] https://github.com/modultechnology/nel_errors

7. Bibliographical References

Brasoveanu, A., Nixon, L. J. B., Weichselbraun, A., and Scharl, A. (2016). A regional news corpora for contextualized entity discovery and linking. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, pages 3333–3338.

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In Daniel Schwabe, et al., editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 249–260. International World Wide Web Conferences Steering Committee / ACM.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity extraction. In Marta Sabou, et al., editors, *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124. ACM.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Hachey, B., Nothman, J., and Radford, W. (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*, pages 464–469. The Association for Computer Linguistics.

Heinzerling, B. and Strube, M. (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*, pages 37–42. ACL.

Hellmann, S., Lehmann, J., Auer, S., and Nitzschke, M. (2012). NIF combinator: Combining NLP tool output. In Annette ten Teije, et al., editors, *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, volume 7603 of *Lecture Notes in Computer Science*, pages 446–449. Springer.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792. ACL.

Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., and Weikum, G. (2012). KORE: keyphrase overlap relatedness for entity disambiguation. In Xue-wen Chen, et al., editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 545–554. ACM.

Jha, K., Röder, M., and Ngomo, A. N. (2017). All that glitters is not gold - rule-based curation of reference datasets for named entity recognition and entity linking. In Eva Blomqvist, et al., 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.

Ji, H. and Nothman, J. (2016). Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end kbp. In *Eighth Text Analysis Conference (TAC)*. NIST.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Radford, W., Nothman, J., Curran, J. R., Hachey, B., and Honnibal, M. (2011). Naïve but effective NIL clustering baselines - CMCRC at TAC 2011. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST.

Radford, W. (2015). *Linking Named Entities to Wikipedia*. Ph.D. thesis, School of Information Technologies, Faculty of Engineering and IT, The University of Sydney.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., and Both, A. (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.

Usbeck, R., Röder, M., Ngomo, A. N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L. (2015). GERBIL: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, pages 1133–1143.

van Erp, M., Mendes, P. N., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., and Waitelonis, J. (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.

Weichselbraun, A., Streiff, D., and Scharl, A. (2015). Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools*, 24(2):1–31.