Posthoc Verification and the Fallibility of the Ground Truth

Yifan Ding, Nicholas Botzer, Tim Weninger
Department of Computer Science & Engineering
University of Notre Dame
Notre Dame, IN, USA
{yding4,nbotzer,tweninge}@nd.edu

Abstract

Classifiers commonly make use of pre-annotated datasets, wherein a model is evaluated by pre-defined metrics on a held-out test set typically made of human-annotated labels. Metrics used in these evaluations are tied to the availability of well-defined ground truth labels, and these metrics typically do not allow for inexact matches. These noisy ground truth labels and strict evaluation metrics may compromise the validity and realism of evaluation results. In the present work, we conduct a systematic label verification experiment on the entity linking (EL) task. Specifically, we ask annotators to verify the correctness of annotations after the fact (*i.e.*, posthoc). Compared to pre-annotation evaluation, state-of-the-art EL models performed extremely well according to the posthoc evaluation methodology. Surprisingly, we find predictions from EL models had a similar or higher verification rate than the ground truth. We conclude with a discussion on these findings and recommendations for future evaluations. The source code, raw results, and evaluation scripts are publicly available via the MIT license at https://github.com/yifding/e2e_EL_evaluate

The general machine learning pipeline starts with a dataset (a collection of documents, images, medical records, etc.). When labels are not inherent to the data, they must be annotated – usually by humans. A label error occurs when an annotator provides a label that is "incorrect." But this raises an interesting question: who gets to decide that some annotation is incorrect?

One solution is to ask k annotators and combine their labels somehow (*e.g.*, majority vote, probability distribution). Subjectivity comes into play here. Given identical instructions and identical items, some annotators may focus on different attributes of the item or have a different interpretation of the labeling criteria. Understanding and modelling label uncertainty remains a compelling challenge in evaluating machine learning systems (Sommerauer, Fokkens, and Vossen, 2020; Resnick et al., 2021).

Figure 1: Example Entity Linking task where the pre-annotated ground truth mention and link is different from the predicted label. Standard evaluation regimes count this as a completely incorrect prediction despite being a reasonable label.

Tasks that require free-form, soft, or multi-class annotations present another dimension to this challenge. For example, natural language processing tasks like named entity recognition (NER) and entity linking (EL) rely heavily on datasets comprised of free-form human annotations. These tasks are typically evaluated against a held out portion of the already-annotated dataset. A problem arises when NER and EL tasks produce labels that are not easily verified as "close enough" to the correct groundtruth (Ribeiro et al., 2020). Instead, like the example in Fig. 1, most NER and EL evaluation metrics require exact matches against free-form annotations (Sevgili et al., 2020; Goel et al., 2021). This strict evaluation methodology may unreasonably count labels that are "close enough" as incorrect and is known to dramatically change performance metrics (Gashteovski et al., 2020).

Producing a *verifiable* answer is not the same as producing the *correct* answer. This distinction is critical. Asking a machine learning system to independently provide the same label as an annotator is a wildly different task than asking an annotator to verify the output of a predictor (*posthoc verification*). Unfortunately the prevailing test and evaluation regime requires predictors to exactly match noisy, free-form, and subjective human annotations. This paradigm represents a mismatch

Proceedings of the First Workshop on Dynamic Adversarial Data Collection, pages 23 - 29
July 14, 2022

Table 1: Statistics of the entity linking datasets and annotations.

	Datasets	Docs	Annotations GT	Annotations E2E	Annotations REL	Tasks GT	Tasks E2E	Tasks REL	Verified Annotations GT	Verified Annotations E2E	Verified Annotations REL
AIDA	**AIDA-train**	946	18541*	18301	21204	2801	2802	2913	18511	18274	21172
	AIDA-A	216	4791	4758	5443	713	715	725	4787	4754	5439
	AIDA-B	231	4485	4375	5086	636	646	654	4480	4370	5079
WNED	**ACE2004**	57*	257	1355	1675	114	318	334	256	1352	1672
	AQUAINT	50	727	810	925	175	170	179	727	810	925
	CLUEWEB	320	11154	12273	23114	3526	3678	4944	11139	12247	23056
	MSNBC	20	656	629	756	164	163	171	656	629	756
	WIKIPEDIA	345*	6793*	8141	11184	1348	1578	1638	6786	8136	11177

* indicate results different from related work because they remove out-of-dictionary annotations.

that, if left unaddressed, threatens to undermine future progress in machine learning.

Main Contributions. We show that the distinction between pre-annotated and posthoc-annotated labels is substantial and the distinction presents consequences for how we determine the state-of-the-art in machine learning systems.

We conducted systematic experiments using posthoc analysis on a large case study of eight popular entity linking datasets with two state-of-the-art entity linking models, and report some surprising findings: First, state-of-the-art EL models generally predicted labels with *higher* verification rate than the ground truth labels. Second, there was substantial disagreement among annotators as to what constitutes a label that is "good enough" to be verified. Third, a large proportion (between 10%-70% depending on the dataset) of verified entities were missing from the ground truth dataset.

The Setting: Entity Linking

The goal of EL is to identify words or phrases that represent real-world entities and match each identified phrase to a listing in some knowledge base. Like most classification systems, EL models are typically trained and tested on large pre-annotated benchmark datasets. Table 1 describes eight such benchmark datasets that are widely used throughout the EL and broader NLP communities.

EL Models. In order to better understand the effect of pre-annotated benchmarks on machine learning systems, it is necessary to test a handful of state-of-the-art EL systems. Specifically, we chose: (1) The end-to-end (E2E) entity linking model, which generates and selects span candidates with associated entity labels. The E2E model is a word-level model that utilizes word and entity embeddings to compute span-level contextual

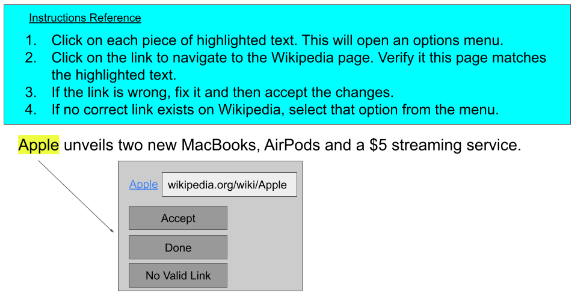

Figure 2: Web system used to collect posthoc annotations from workers.

scores. Word and entity embeddings are trained on Wikipedia, and the final model is trained and validated using AIDA-train and AIDA-A respectively (Kolitsas, Ganea, and Hofmann, 2018). (2) The Radboud Entity Linker (REL), which combines the Flair (Akbik, Blythe, and Vollgraf, 2018) NER system with the mulrel-nel (Le and Titov, 2018) entity disambiguation system to create a holistic EL pipeline (van Hulst et al., 2020). In addition, our methodology permits the evaluation of the GT as if it were a competing model. The relative performance of E2E and REL can then compared with the GT to better understand the performance of the posthoc annotations.

Data collection. We have previously argued that these evaluation metrics may not faithfully simulate *in vivo* performance because (1) the ground truth annotations are noisy and subjective, and (2) exact matching is too strict. We test this argument by collecting posthoc verifications of the three models, including the pre-annotated GT, over the datasets.

We created a simple verification system, illustrated in Fig. 2, and used Amazon Mechanical Turk to solicit workers. For each document and model, we asked a single worker to verify all present entity annotations (*i.e.*, an entity mention and its linked entity). Annotators can then choose to (1) Verify

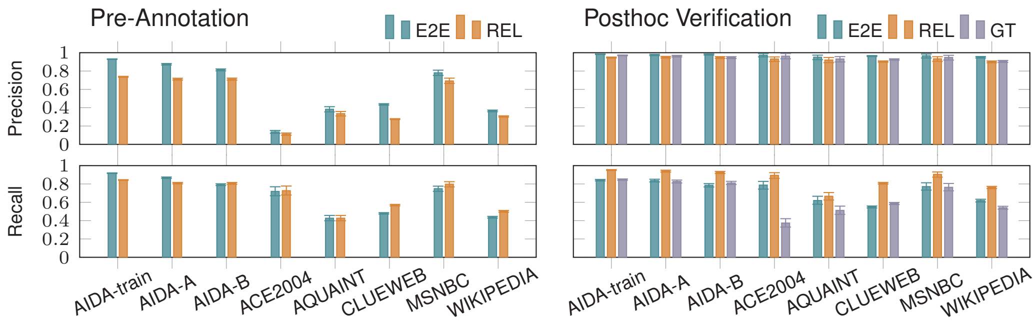

Figure 3: Precision and recall results from pre-annotation evaluation (Left) compared with the posthoc verification evaluation (Right). Error bars represent 95% confidence intervals on bootstrapped samples of the data. Posthoc verification returns substantially higher scores than the pre-annotation evaluation.

the annotation (2) Modify the annotation, or (3) Remove the annotation.

- **Verify**: The annotator determines that the current annotation (both mention and Wikipedia link) is appropriate.
- **Modify**: The annotator determines that the Wikipedia link is incorrect. In this case, they are asked to search and select a more appropriate Wikipedia link, use it to replace the existing link, and then accept the new annotation.
- **Remove**: The annotator determines that the current mention (highlighted text) is not a linkable entity. In this case, they remove the link from the mention.

We made a deliberate decision to not permit new annotation of missing entity mentions. That is, if the model did not label an entity, then there is no opportunity for the worker to add a new label. This design decision kept the worker focused on the verification task, but possibly limits the coverage of the verified dataset. We provide further comments on this decision in the Results section.

Each annotator is assigned to 20 tasks including one control task with three control annotations. We only accept and collect annotations from workers that passed the control task.

We paid each worker 3 USD for each HIT. We estimate a average hourly rate of about 9 USD; and paid a total of 6,520 USD. From these, we received 167,432 annotations. The breakdown of tasks, annotations shown to workers, and verified annotations are listed in Table 1 for each dataset and model.

Prior to launch, this experiment was reviewed and approved by an impaneled ethics review board at the University of Notre Dame. The source code, raw results, and evaluation scripts are publicly available via the MIT license at https://github.com/yifding/e2e_EL_evaluate

Posthoc Verification Methodology

The Pre-Annotation Evaluation Regime. First, we re-tested the E2E and REL models and evaluated their micro precision and recall under the typical pre-annotation evaluation regime. These results are illustrated in Fig 3 and are nearly identical to those reported by related works (Kolitsas, Ganea, and Hofmann, 2018; van Hulst et al., 2020).

Posthoc Verification Evaluation

Our next task is to define appropriate evaluation metrics that can be used to compare the results of the posthoc verification experiment with results from the pre-annotation evaluation regime.

Verification Rate. For each combination of dataset and model providing annotations, we compute the verification rate as the percentage of annotations that were verified. Formally, let $d \in$ datasets; $m \in$ models; and $V_{m,d}$ be the set of verified annotations in a pairing of d and m Likewise, let $N_{d,m}$ be the pre-annotations of model m on dataset d. We therefore define the verification rate of a dataset-model pair as $r_{m,d} = |V_{m,d}|/|N_{d,m}|$. Higher verification rates indicate that the dataset contains annotations and/or the model is more capable of providing labels that pass human inspection.

Verification Union. It is important to note that each model and document was evaluated by only a single worker. However, we were careful to assign each worker annotations randomly drawn from model/document combinations. This randomization largely eliminates biases in favor or against any model or dataset. Furthermore, this methodology provides for repetitions when annotations match exactly across models – which is what models are optimized for in the first place! In this scenario the union of all non-exact, non-overlapping annotations provides a superset of annotations similar to how pooling is used in information retrieval evaluation to create a robust result set (Zobel, 1998). Formally, we define the verification union of a dataset d as $V_d = \bigcup_m V_{m,d}$.

Posthoc Precision and Recall. The precision metric is defined as the ratio of true predictions to all predictions. If we recast the concept of true predictions to be the set of verified annotations $V_{m,d}$, then it is natural to further consider $N_{d,m}$ to be the set of all predictions for some dataset and model pair, especially considering our data collection methodology restricts $V_{m,d} \subseteq N_{d,m}$. Thus the posthoc precision of a model-data pairing is simply the verification rate $r_{m,d}$.

The recall metric is defined as the ratio of true predictions to all true labels. If we keep the recasting of true positives as verified annotations $V_{m,d}$, then all that remains a definition of true labels. Like in most evaluation regimes the set of all true labels is estimated by the available labels in the dataset. Here, we do the same and estimate the set of true labels as the union of a dataset's verified annotations V_d. Thus posthoc recall of a model-data pairing is $|V_{m,d}|/|V_d|$.

Posthoc Verification Results

Using the evaluation tools introduced in the previous section, we begin to answer interesting research questions. First, do the differences between evaluation regimes, *i.e.*, pre-annotation versus posthoc verification, have any affect on our perception of model performance.

To shed some light on this question, we compared the precision and recall metrics calculated using the pre-annotation evaluation regime against the precision and recall metrics calculated using the posthoc verification regime. The left quadplot in Fig. 3 compares model performance under the different evaluation regimes. Error bars represent the empirical 95% confidence internals drawn from 1000 bootstrap samples of the data. We make two major conclusions from this comparison:

Pre-annotation performance is lower than Posthoc verification. The differences between the scores of the pre-annotation compared to posthoc verification are striking. Posthoc annotation shows very good precision scores across all datasets. Although the models may not exactly predict the pre-annotated label, high posthoc precision indicates that their results appear to be "close-enough" to obtain human verification.

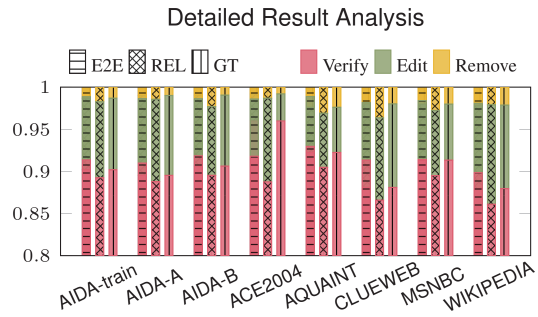

Figure 4: Detailed error analysis of verification rates in Fig. 3(top right). The E2E model consistently outperforms the ground truth (GT).

Conclusion: the widely-used exact matching evaluation regime is too strict. Despite its intention, the pre-annotation evaluation regime does not appear to faithfully simulate a human use case.

Machine Learning models outperform the Ground Truth. The posthoc verification methodology permits the GT annotations to be treated like any other model, and are therefore included in Fig. 3 (right plot). These results were unexpected and surprising. We found that labels produced by the EL models oftentimes had a higher verification rate than the pre-annotated ground truth. The recall metric also showed that the EL models were also able to identify more verified labels than GT.

Conclusion: Higher precision performance of the EL models indicates that human annotators make more unverifiable annotations than the EL models. Higher recall performance of the EL models also indicates that the EL models find a greater coverage of possible entities. The recall results are less surprising because human annotators may be unmotivated or inattentive during free-form annotation – qualities that tend to not affect EL models.

Error Analysis of the Ground Truth

For each linked entity, the posthoc verification methodology permitted one of three outcomes: verification, modification, or removal. The plot in Fig. 4 shows the percentage of each outcome for each model and dataset pair; it is essentially a zoomed-in, more-detailed illustration of the Posthoc Verification Precision result panel from Fig. 3, but with colors representing outcomes and patterns representing models. Edits indicate that the named entity recognition (*i.e.*, mention detection) portion of the EL model was able to identify an entity, but the entity was not linked to a verifiable entity. The available dataset has an enumeration of corrected linkages, but we do not consider them further in the present work. Removal indicates an error with the mention detection. From these results we find that, when a entity mention is detected it is usually a good detection; the majority of the error comes from the linking subtask.

A similar error analysis of missing entities is not permitted from the data collection methodology because we only ask workers to verify pre-annotated or predicted entities, not add missing entities. Because all detected mentions are provided with some entity link, we can safely assume that missing entities is mostly (perhaps wholly) due to errors in the mention detection portion of EL models.

Discussion

The primary goal of the present work is to compare pre-annotation labels contributed by human workers against verified annotations of the same data. Using entity linking as an example task, we ultimately found that these two methodologies returned vastly different performance results. From this observation we can draw several important conclusions. First, EL models have a much higher precision than related work reports. This difference is because the standard evaluation methodology used in EL, and throughout ML generally, do not account for soft matches or the semantics of what constitutes a label that is "close enough". Our second conclusion is that EL models, and perhaps ML models generally, sometimes perform better than ground truth annotators – at least, that is, according to other ground truth annotators.

Acknowledgments

This research is sponsored in part by the Defense Advanced Research Projects Agency (DAPRA) under contract numbers HR00111990114 and HR001121C0168. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Aharoni, R.; and Goldberg, Y. 2018. Split and Rephrase: Better Evaluation and Stronger Baselines. In ACL, 719–724.

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In COLING, 1638–1649.

Belinkov, Y.; and Bisk, Y. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In ICLR.

Botzer, N.; Ding, Y.; and Weninger, T. 2021. Reddit entity linking dataset. Information Processing & Management, 58(3): 102479.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Bowman, S. R.; and Dahl, G. E. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? arXiv preprint arXiv:2104.02145.

Chzhen, E.; Denis, C.; Hebiri, M.; and Lorieul, T. 2021. Set-valued classification–overview via a unified framework. arXiv preprint arXiv:2102.12318.

Denis, C.; and Hebiri, M. 2017. Confidence sets with expected sizes for multiclass classification. JMLR, 18(1): 3571–3598.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL HLT, 4171–4186.

Ganea, O.-E.; and Hofmann, T. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In EMNLP, 2619–2629.

Gashteovski, K.; Gemulla, R.; Kotnis, B.; Hertling, S.; and Meilicke, C. 2020. On Aligning OpenIE Extractions with Knowledge Bases: A Case Study. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 143–154.

Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In EMNLP, 1161–1166.

Glockner, M.; Shwartz, V.; and Goldberg, Y. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In ACL, 650–655.

Goel, K.; Rajani, N.; Vig, J.; Tan, S.; Wu, J.; Zheng, S.; Xiong, C.; Bansal, M.; and Ré, C. 2021. Robustness Gym: Unifying the NLP Evaluation Landscape. arXiv preprint arXiv:2101.04840.

Graham, Y.; Baldwin, T.; Moffat, A.; and Zobel, J. 2013. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 33–41.

Graham, Y.; Baldwin, T.; Moffat, A.; and Zobel, J. 2014. Is machine translation getting better over time? In EACL, 443–451.

Guo, Z.; and Barbosa, D. 2014. Robust entity linking via random walks. In CIKM, 499–508.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324.

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 EMNLP, 782–792.

Hynes, N.; Sculley, D.; and Terry, M. 2017. The data linter: Lightweight, automated sanity checking for ml data sets. In NIPS MLSys Workshop.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; et al. 2021. Dynabench: Rethinking Benchmarking in NLP. arXiv preprint arXiv:2104.14337.

Kolitsas, N.; Ganea, O.-E.; and Hofmann, T. 2018. End-to-end neural entity linking. CoNLL.

Le, P.; and Titov, I. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In ACL, 1595–1604.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 3: 211–225.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, 74–81.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. NeurIPS, 30: 4765–4774.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In ACL, 142–150.

Northcutt, C. G.; Athalye, A.; and Mueller, J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.

Oortwijn, Y.; Ossenkoppele, T.; and Betti, A. 2021. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), 131–141.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL, 311–318.

Park, C. W.; Jun, S. Y.; and MacInnis, D. J. 2000. Choosing what I want versus rejecting what I do not want: An application of decision framing to product option choice decisions. Journal of Marketing Research, 37(2): 187–202.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing bayesian models of annotation. TACL, 6: 571–585.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis Only Baselines in Natural Language Inference. NAACL HLT, 180.

Prabhakaran, V.; Hutchinson, B.; and Mitchell, M. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In EMNLP, 5744–5749.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In ACL, 784–789.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Resnick, P.; Kong, Y.; Schoenebeck, G.; and Weninger, T. 2021. Survey Equivalence: A Procedure for Measuring Classifier Accuracy Against Human Labels. arXiv preprint arXiv:2106.01254.

Ribeiro, M. T.; Guestrin, C.; and Singh, S. 2019. Are red roses red? evaluating consistency of question-answering models. In ACL, 6174–6184.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In ACL, 4902–4912.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In EMNLP, 193–203.

Rolnick, D.; Veit, A.; Belongie, S.; and Shavit, N. 2017. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694.

Rosales-Méndez, H.; Hogan, A.; and Poblete, B. 2019a. Fine-grained evaluation for entity linking. In EMNLP-IJCNLP, 718–727.

Rosales-Méndez, H.; Hogan, A.; and Poblete, B. 2019b. NIFify: Towards Better Quality Entity Linking Datasets. In WWW 2019, 815–818.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In CHI, 1–15.

Schwartz, R.; Sap, M.; Konstas, I.; Zilles, L.; Choi, Y.; and Smith, N. A. 2017. Story cloze task: Uw nlp system. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, 52–55.

Sevgili, O.; Shelmanov, A.; Arkhipov, M.; Panchenko, A.; and Biemann, C. 2020. Neural entity linking: A survey of models based on deep learning. arXiv preprint arXiv:2006.00575.

Sommerauer, P.; Fokkens, A.; and Vossen, P. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In COLING, 4798–4809.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2020. Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2016. Newsqa: A machine comprehension dataset. arXiv preprint arXiv:1611.09830.

Tsuchiya, M. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In LREC.

van Hulst, J. M.; Hasibi, F.; Dercksen, K.; Balog, K.; and de Vries, A. P. 2020. REL: An entity linker standing on the shoulders of giants. In SIGIR, 2197–2200.

Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In SIGIR, 307–314.