

Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards

Varvara Arzt

Faculty of Informatics, TU Wien
D!ARC, University of Klagenfurt
varvara.arzt@tuwien.ac.at

Allan Hanbury

Faculty of Informatics, TU Wien
allan.hanbury@tuwien.ac.at

Abstract

This paper investigates the transparency in the creation of benchmarks and the use of leaderboards for measuring progress in NLP, with a focus on the relation extraction (RE) task. Existing RE benchmarks often suffer from insufficient documentation, lacking crucial details such as data sources, inter-annotator agreement, the algorithms used for the selection of instances for datasets, and information on potential biases like dataset imbalance. Progress in RE is frequently measured by leaderboards that rank systems based on evaluation methods, typically limited to aggregate metrics like F1-score. However, the absence of detailed performance analysis beyond these metrics can obscure the true generalisation capabilities of models. Our analysis reveals that widely used RE benchmarks, such as TACRED and NYT, tend to be highly imbalanced and contain noisy labels. Moreover, the lack of class-based performance metrics fails to accurately reflect model performance across datasets with a large number of relation types. These limitations should be carefully considered when reporting progress in RE. While our discussion centers on the transparency of RE benchmarks and leaderboards, the observations we discuss are broadly applicable to other NLP tasks as well. Rather than undermining the significance and value of existing RE benchmarks and the development of new models, this paper advocates for improved documentation and more rigorous evaluation to advance the field.

1 Introduction

We examine the transparency in benchmarks and leaderboards, focusing on the relation extraction (RE) task. Our analysis utilises two broadly accepted RE datasets, TACRED (Zhang et al., 2017) and NYT (Riedel et al., 2010). While this paper focuses on the transparency of RE benchmarks and leaderboards, the observations we discuss are also

relevant to other areas of natural language processing (NLP).

The development of state-of-the-art (SOTA) models in NLP is heavily reliant on benchmarks for evaluation. These benchmarks not only serve as a standard for assessing model performance but also play a pivotal role in shaping the perceived progress within the field. However, the current benchmarks often lack transparency in regard to their creation process, which can significantly impact the reliability of the evaluations conducted using them.

Opaque benchmarks and the absence of detailed performance analysis can obscure the true generalisation capabilities of models (Geburu et al., 2021; Dehghani et al., 2021). When benchmarks are not fully transparent — lacking comprehensive metadata, clear articulation of limitations, and rigorous evaluation reports — their ability to accurately reflect a model’s robustness and generalisability is compromised. This can lead to an overestimation of progress, as models may appear to perform well on certain benchmarks but fail to generalise effectively to different or more complex datasets.

To enhance transparency and reproducibility in the evaluation of models, it is essential to publish the annotation guidelines and instructions that were provided to annotators during the creation of benchmarks. Understanding the exact criteria and procedures used in annotation is critical for interpreting the results obtained from these benchmarks and for comparing the performance of different models.

It is also important to recognise that widely used benchmarks such as TACRED (Zhang et al., 2017), TACRED-RE (Alt et al., 2020), and NYT (Riedel et al., 2010) cover only a subset of all possible relations. This limitation should be considered when evaluating models, as these benchmarks do not necessarily capture the full complexity of relation extraction task.

Additionally, when asserting that a new system

outperforms existing ones, it is crucial to provide more granular results beyond aggregate metrics like weighted average or macro F1-score. These metrics, while useful, may not be sufficiently informative, particularly in the context of datasets with a large number of labels (Dehghani et al., 2021) and significant class imbalances, such as NYT or TACRED.

Although this position paper addresses certain issues with existing RE benchmarks and evaluation approaches, it does not seek to diminish their significance or the value of developing new RE models, which are crucial for advancing the NLP field. Instead, it aims to promote improved documentation of benchmarks and the adoption of more rigorous evaluation practices for SOTA RE systems.

2 Related Work

Despite the critical role of data in NLP, the documentation of the creation process of existing datasets remains scarce, unstandardised, and often lacks transparency, even for publicly available datasets (Bender and Friedman, 2018; Gebru et al., 2021; Peng et al., 2021; Singh, 2023; Kovatchev and Lease, 2024).

Gebru et al., 2021 addresses the issue of insufficient transparency in dataset creation by proposing that dataset creators accompany each dataset with a datasheet. This datasheet would document essential information about the dataset’s creation process, thereby enhancing the reproducibility of machine learning experiments and helping to mitigate potential biases. They outline seven key stages of the dataset lifecycle: motivation, composition, collection process, preprocessing/cleaning/labeling, intended uses, distribution, and maintenance.

The lack of transparency in benchmark creation significantly impacts the evaluation of models trained on these benchmarks. As Kovatchev and Lease, 2024 highlights, many evaluation frameworks operate under the implicit assumption that a particular dataset is representative of the task it is intended to benchmark. However, systematic approaches to testing model generalisation remain limited (Hupkes et al., 2023). To address this gap, Kovatchev and Lease, 2024 propose the use of dataset similarity vectors, which consider various dimensions of the data, such as noise and ambiguity features, to more accurately predict the generalisation capabilities of models trained on these datasets.

Hupkes et al., 2023 present a comprehensive taxonomy of methods for studying the generalisation capabilities of models and introduce the GenBench evaluation card template¹ to assist researchers in systematically documenting, justifying, and tracing their generalisation experiments. Evaluating the generalisation capabilities of models has become increasingly complex in the era of large language models (LLMs), which strive to achieve human-like generalisation but are trained on vast, uncontrolled, and often nontransparent datasets.

To enhance the transparency of model evaluation processes, researchers advocate for testing new SOTA models in challenging scenarios involving perturbed instances (Wu et al., 2019; Gardner et al., 2020; Goel et al., 2021), thereby assessing model capabilities in more realistic settings than those provided by traditional test sets. Linzen, 2020, when discussing the limitations of current evaluation approaches, particularly in the context of developing systems with human-like generalisation capabilities, introduces the Generalisation Leaderboards. These leaderboards evaluate systems on test sets derived from distributions different from those used during training. This approach addresses the limitation that testing a model on data drawn from the same distribution as the training set does not necessarily demonstrate the model’s ability to effectively solve the task; rather, it may merely reflect the model’s proficiency in capturing statistical patterns specific to the training data.

In addition to traditional leaderboards, which often rank SOTA systems based solely on holistic metrics such as aggregate F1-score, Liu et al., 2021 propose leaderboards that incorporate more fine-grained metrics and offer functionality for direct analysis of misclassifications. This approach allows for a more detailed comparison of system performance, enabling users to directly identify the strengths and weaknesses of specific systems, thereby enhancing the transparency of leaderboards.

3 Transparency in Benchmark Creation

Current relation extraction benchmarks still lack transparency in their creation processes, making it difficult to assert that they generalise well on out-of-distribution data. For instance, we often lack detailed information about the text sources used to create these benchmarks. Transparency in the cre-

¹Available at https://genbench.org/eval_cards/.

ation of RE datasets is crucial not only for mitigating potential biases but also for facilitating progress in the field. By better understanding the limitations of existing RE benchmarks we are able to consequently better understand the limitations of systems that make use of these data. We examine the problem of lacking transparency through the lens of two of the most widely used general-purpose relation extraction benchmarks, namely NYT (Riedel et al., 2010) and TACRED (Zhang et al., 2017) datasets. These benchmarks are broadly accepted by the NLP community and continue to be widely used, even in the era of LLMs (Huguet Cabot and Navigli, 2021; Wang et al., 2021; Tang et al., 2022; Wang et al., 2022; Efeoglu and Paschke, 2024; Sainz et al., 2024). Both the NYT and TACRED datasets address the task of sentence-level relation extraction.

3.1 Analysis of NYT and TACRED Datasets: Transparency and Limitations

The NYT dataset contains 24 relation types as well as a ‘None’ class and is based on a corpus of New York Times newspaper articles (Riedel et al., 2010). As Table 1 shows, the dataset includes over 266k sentences, with 64% of the instances belonging to the ‘None’ class.

Table 1: NYT Dataset

Type	Number of Samples
Positive Samples	96,228
Negative Samples	170,021
Total	266,249

NYT is created through distant supervision, utilising corpus of the New York Times articles (Sandhaus, 2008) and using Freebase (Bollacker et al., 2008) as the external supervision source. Detailed information on the included relation types and the number of instances for each relation can be found in Table 4 in the Appendix. The NYT dataset is publicly available. The example in Figure 1 shows one of the instances from the NYT dataset, which illustrates the issues associated with using distant supervision for dataset creation.²

As illustrated by the NYT instance in Figure 1, the sentence is labeled as containing the relation

²The example in Figure 1 represents NYT instance with article ID ‘/m/vinci8/data1/riedel/projects/relation/kb/nyt1/docstore/nyt-2005-2006.backup/1677367.xml.pb’. The NYT dataset can be found at <https://github.com/INK-USC/ReQuest>.

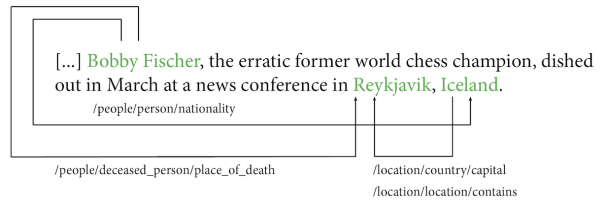


Figure 1: Example from the NYT dataset

‘/people/person/nationality’ between the head entity Bobby Fischer and the tail entity Iceland. However, this relation is not directly mentioned in text. This issue arises from the distant supervision method used to create the NYT dataset: when named entities are connected by a ‘nationality’ relation in Freebase, it though does not necessarily mean that this relation is explicitly present in the NYT data. Such interpretations can introduce significant biases in relation extraction systems and do not, for instance, reliably demonstrate a system’s ability to detect ‘nationality’ relation in general. Such a reasoning pattern can be questioned as valid and would probably be labeled as a hallucination in the era of LLMs. The problem of noise in relation extraction datasets created using distant supervision has been discussed in several works, including Yaghoobzadeh et al., 2017.

The TACRED dataset contains 41 relations as well as a ‘no_relation’ class. TACRED includes over 106k instances, though, as shown in Table 2, 80% of the instances belong to ‘no_relation’ class, making the dataset highly imbalanced.

Table 2: TACRED Dataset

Type	Number of Samples
Positive Samples	21,773
Negative Samples	84,491
Total	106,264

The TACRED dataset is a fully supervised dataset obtained via crowdsourcing, and is based on the TAC KBP³ corpus, which includes English newswire and web text. It is distributed under the Linguistic Data Consortium (LDC) license. Detailed information on the included relation types and the number of instances for each relation in TACRED can be found in Table 3 in the Appendix. The example in Figure 2 shows one of the instances from the TACRED dataset.⁴

³<https://tac.nist.gov/2017/KBP/index.html>

⁴The example in Figure 2 originates from the paper de-

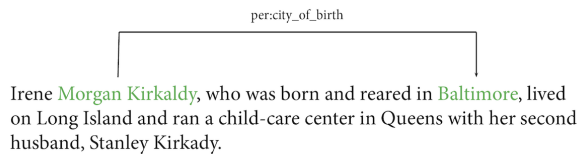


Figure 2: TACRED example

Compared to the NYT example in Figure 1, where the relation type cannot be determined solely from the information provided in the sentence, the TACRED instance in Figure 2 demonstrates an explicit relation that can be directly extracted from the sentence without requiring additional, potentially biased, reasoning steps. However, the TACRED dataset restricts each sentence to contain only one relation, in contrast to the NYT dataset, which allows each sentence to have multiple labels. The formulation of the relation extraction task like in TACRED can lead to many false negatives (Xie et al., 2021). For instance, the example above contains relations like ‘per:stateorprovinces_of_residence’, and ‘per:employee_of’ or ‘per:spouse’, all of which are part of the TACRED list of relations. Thus, restricting each instance to a single relation is far removed from the complexity of real-world text and may significantly mislead the model.

Despite the supplementary material⁵ provided additionally to the paper (Zhang et al., 2017), TACRED still lacks transparency regarding the limitations mentioned above, as well as clarity on how the instances for the dataset were selected from the TAC KBP corpus. A similar lack of transparency exists with the NYT dataset, where the selection process for instances from the NYT corpus (Sandhaus, 2008) is not clearly documented. Since access to the TAC KBP corpus is restricted and the selection process from the NYT corpus is unclear, analysing the data included in both datasets and estimating the generalisation capabilities of models trained on these datasets becomes even more challenging.

Additionally, although TACRED instances were manually annotated by crowd workers, unlike the

describing TACRED (Zhang et al., 2017).

⁵Zhang et al., 2017 mention supplementary material to the main paper describing the dataset, though they do not provide the direct link to material. It can be assumed they were referring to information available at <https://nlp.stanford.edu/projects/tacred/> or <https://tac.nist.gov/2017/KBP/ColdStart/guidelines.html>, but these sources still lack precise details on e.g. the data collection process and the creation of the relations inventory.

NYT instances, the crowdsourced annotations may still be quite noisy. For instance, Alt et al., 2020 demonstrated that over 50% of the challenging ‘no_relation’ instances in the development and test sets of TACRED were mislabeled.

As shown in Tables 1 and 2, both NYT and TACRED are highly imbalanced. While it may appear that the NYT dataset (with 64% of the instances belonging to the ‘None’ class) is more balanced compared to TACRED (where 80% of the instances belong to ‘no_relation’ class), a closer examination on the number of instances for each of the 24 relation types, as detailed in Table 4 in the Appendix, reveals a different picture. Nearly half of the positive instances in the NYT dataset belong to a single relation type, ‘/location/location/contains’, and six out of 24 relations are represented by fewer than 50 instances. For instance, the relation ‘/people/person/profession’ contains only two instances, and ‘/business/company/industry’ has just one instance.

In addition to providing precise information on the data selection process, it is also important to make the annotation guidelines publicly available if human annotators were involved, or to describe the algorithms used if a dataset was created through distant supervision (e.g., prompts). The publication of annotation guidelines and dataset description in general is among others crucial for clarifying ambiguous relations, whose scope may be understood in multiple ways, such as ‘per:title’ in TACRED or ‘None’ in NYT. The ‘None’ label in NYT could indicate either that none of the 24 specified relations apply — implying another, unspecified relation — or that there is no relation at all between the entities. These are fundamentally different scenarios, and conflating them could lead to significant confusion in the model.

Despite the publicly available TAC KBP guidelines⁶, it remains unclear whether this version of annotation guidelines was also provided to the TACRED crowd workers. Furthermore, it is still unclear how the annotators of the TACRED dataset were instructed to handle sentences that contained a relation not listed among the 41 relations, or how they were to deal with sentences containing multiple different relations, as in Figure 2 above.

To introduce clarity in benchmark creation process, it is therefore crucial to publish not only anno-

⁶https://tac.nist.gov/2014/KBP/ColdStart/guidelines/TAC_KBP_2014_Slot_Descriptions_V1.4.pdf

tation guidelines but also the instructions provided to the annotators. While [Riedel et al., 2010](#) describes the process of creating the NYT dataset in a relatively detailed way, when they mention the use of human annotators to evaluate a fixed number of extracted relations in a distant supervision scenario, they still do not provide details on how these human annotators were instructed.

3.2 The Need for Standardised Benchmark Documentation

The analysis of widely used NYT and TACRED RE benchmarks, along with their available documentation, underscores the persistent issue of lacking exhaustive documentation regarding the creation processes of NLP benchmarks. Proper documentation should be easily discoverable and ideally stored according to accepted standards. Currently, information on NLP benchmarks is dispersed across many resources and often lacks the necessary details to make the benchmark creation process fully transparent which is among others crucial for the analysis of generalisation capabilities of a particular dataset. While sources like PaperswithCode⁷ are helpful, they still miss a significant amount of information needed to achieve this goal.

[Gebru et al., 2021](#) addresses the issue of insufficient benchmark transparency and suggests that each new benchmark should be accompanied by a datasheet. The suggested datasheet would include information such as potential sources of noise and errors in the dataset, to enhance transparency and allow for more accurate assessments of the dataset reliability and generalisation capabilities.

The NLP community would greatly benefit from a standardised approach to benchmark documentation, similar to model cards for model reporting ([Mitchell et al., 2019](#)), but specifically designed for datasets. This is at least as important as model metadata. Model cards, which are essentially files containing metadata with useful information about a model in question, have proven effective, as seen in their implementation at HuggingFace.⁸ A similar concept for datasets would ensure that critical information about benchmark creation, potential biases, and other relevant details are systematically recorded and easily accessible. While HuggingFace provides dataset cards⁹ ([Park and Jeong,](#)

[2022](#)) which are a promising step in this direction, most datasets shared via HuggingFace currently have only a fraction of the possible metadata filled out. Moreover, while the ecosystem that HuggingFace provides has undoubtedly contributed significantly to the NLP community, it is essential to acknowledge that, given the open nature of the platform where anyone can upload models and datasets, the reliability of sources, including datasets and their associated metadata, should be approached with caution. Ideally, comprehensive documentation of benchmarks should originate directly from their creators.

A datasheet for benchmarks would ideally include properties such as descriptive and social impact metadata ([Park and Jeong, 2022](#)) including data provenance, data preprocessing details (e.g., filtering approach used to obtain relevant samples), annotation guidelines and other instructions, dataset size, recommended data split information, a list of labels, the specific task being addressed, the method used for creating the benchmark (e.g., human annotation or distant supervision), inter-annotator agreement (if human annotators were involved), and potential sources of noise (e.g., representativeness of the data). In addition to the proposed datasheets for datasets by [Gebru et al., 2021](#), inspiration can be drawn from dataset templates¹⁰ available in the Open Research Knowledge Graph ([Jaradeh et al., 2019](#)). Although these templates are currently used infrequently by benchmark creators, and often only a small fraction of the possible properties are filled out, their wider adoption could significantly enhance benchmark transparency. For instance, a centralised, standardised approach to documenting benchmarks could help establish a more universal system of labels, making it easier to compare benchmarks within a particular domain: e.g., in the case of RE benchmarks, a standardised set of relations could simplify comparisons across different datasets and models.

The way we document the benchmark creation process is becoming increasingly critical in the era of LLMs, especially as we strive to develop Artificial General Intelligence (AGI) systems with human-like reasoning capabilities ([Chollet, 2019](#); [Hendrycks et al., 2021](#)). As we exhaust real-world data, and with the uncertainty about whether data

⁷<https://paperswithcode.com/>

⁸More on model cards at HuggingFace can be found at <https://huggingface.co/docs/hub/en/model-cards>

⁹<https://huggingface.co/docs/hub/en/>

datasets-cards

¹⁰For instance, the <https://orkg.org/template/R178304> dataset template contains 22 properties like inter-annotator agreement or data availability.

presented as human annotations were truly annotated by humans or generated through LLM prompting, ensuring transparent and thorough documentation is essential for accurately evaluating the systems based on these benchmarks.

4 Transparency in Leaderboard Performance Evaluation

The transparency in the benchmark creation process has a direct impact on the ability to adequately evaluate the system trained on a dataset in question and therefore analyse its generalisation capabilities. One of the ways of measuring progress in particular NLP field are leaderboards. Despite the fact that leaderboards push the NLP field forward, they also lack transparency on evaluation process and mostly are limited to the ranking based on holistic metrics such as accuracy or F1-score (Liu et al., 2021). For instance, both TACRED¹¹ and NYT¹² leaderboards on a widely used platform PaperswithCode rely on F1-score as a holistic metric to rank the RE models.

Moreover, not only traditional leaderboards lack fine-grained metrics in their ranking approach, but also the papers that report SOTA results follow this trend, which can lead to an emphasis on achieving top leaderboard positions rather than genuinely addressing the underlying task — a phenomenon known as SOTA-chasing (Rodriguez et al., 2021).

Recent papers reporting SOTA-performance on NYT, TACRED, and TACRED-RE¹³ (Huguet Cabot and Navigli, 2021; Wang et al., 2022; Tang et al., 2022; Efeoglu and Paschke, 2024; Sainz et al., 2024; Orlando et al., 2024) report only aggregate metrics such as micro or macro f1-score, recall, and precision. Consequently, these evaluations lack more fine-grained, class-based metrics, which are crucial for the analysis of RE systems dealing with a large number of relations such as the 42 labels in TACRED and the 25 labels in NYT. In the context of imbalanced datasets like TACRED and NYT, a system may achieve high overall metrics by always predicting a ‘no_relation’ class. However, this outcome does not indicate that the system has indeed effectively learned to

solve the relation extraction task across a diverse set of over 20 labels. Notably, even the original papers presenting TACRED (Zhang et al., 2017), TACRED-RE (Alt et al., 2020), NYT (Riedel et al., 2010) do not contain fine-grained, class-based metrics. Given the significant class imbalance reflected in the Tables 1 and 2, as well as the fact that many relations in both TACRED and NYT are represented by only a few instances, such as ‘/people/person/profession’ in NYT (see Table 4 in the Appendix), which contains only two instances, reporting class-based metrics is essential for adequately assessing the capabilities of a particular system to solve the RE task. Without detailed performance reports, it is difficult to determine whether a new SOTA system generalises well or simply creates the illusion of improvement through SOTA-chasing.

Benchmarks such as NYT, TACRED, and TACRED-RE lack standardised guidelines for reporting results, leading to inconsistencies across publications that report SOTA results on these benchmarks (Dehghani et al., 2021). This lack of agreement can cause discrepancies in leaderboard rankings. For example, there is no consensus on which aggregated score should be used on platforms like PaperswithCode. The current top-performing model on the TACRED benchmark (Efeoglu and Paschke, 2024) reports the micro-F1 score, which is also used for ranking. In contrast, the current second (Wang et al., 2022) and third (Huang et al., 2022) top-ranked models on TACRED report the macro-F1 score, which is also utilised for their ranking on PaperswithCode. This inconsistency in evaluation metrics raises concerns about the reliability of the leaderboard rankings.

Additionally, there is no overlap between the top-performing models listed on PaperswithCode leaderboards for NYT and TACRED, meaning that all top-performing models for TACRED are different from those for NYT. This further complicates the analysis of these models’ generalisation capabilities and makes it difficult to assess model ranking consistency across RE benchmarks. Focusing exclusively on achieving high performance on a single benchmark, without considering results across multiple benchmarks, can result in models that are overly specialised for specific benchmarks. This, however, does not necessarily indicate meaningful progress in addressing a particular NLP task (Dehghani et al., 2021), such as relation extraction.

¹¹<https://paperswithcode.com/sota/relation-extraction-on-tacred>

¹²<https://paperswithcode.com/sota/relation-extraction-on-nyt>

¹³TACRED-RE is a revised version of original TACRED, with a subset of challenging development and test set instances relabeled by professional annotators (Alt et al., 2020).

Papers reporting SOTA results on RE, including the original TACRED (Zhang et al., 2017), TACRED-RE (Alt et al., 2020), NYT (Riedel et al., 2010), often do not provide information on whether the issue of class imbalance was addressed. Such details should be included in system description papers, particularly when reporting new SOTA results. For instance, the authors of the Biographical RE dataset (Plum et al., 2022) tackled the problem of large class imbalance by removing some of majority class relations, thereby equalising them with the sum of all other relations.

Model performance ceiling (Alt et al., 2020) may be caused by the presence of noisy data, which can limit the potential for improvement by new RE methods. As discussed in Section 3, this noise can originate from both distantly-supervised datasets, such as NYT, and fully-supervised crowdsourced datasets, such as TACRED. Additionally, the way the task is formulated, whether as a single-label (TACRED) or multi-label (NYT) classification task, can contribute to performance limitations. For example, an RE model might make a correct prediction, but due to the task being framed as a single-label classification problem — despite real-world instances potentially containing multiple relations — this could lead to misclassification. Such factors should be considered when reporting new SOTA results. Moreover, in the era of LLMs, it is possible that multiple outputs generated by an LLM for an RE task could be correct (Hendrycks et al., 2021), a nuance that is not captured by holistic metrics like aggregate F1-score.

Due to the mentioned limitations of traditional leaderboards such as the ones utilised on the PaperswithCode platform, Liu et al., 2021 suggest an ExplainaBoard interactive tool that provides both holistic and fine-grained metrics as well as functionality for direct analysis of misclassifications. Such an extension of traditional leaderboards enables the direct detection of strengths and weaknesses of a particular system, as well as of a benchmark, thereby enhancing the ability to assess the generalisation capabilities of systems, such as those used for relation extraction.

Moreover, evaluating model performance on a test set drawn from the same distribution as the training set does not necessarily demonstrate a model’s ability to solve an underlying task (Linzen, 2020), such as relation extraction. To address this issue, Linzen, 2020 propose Generalisation Leaderboards, which would evaluate systems on test sets

derived from different distributions than the training set. For instance, it would be valuable to assess a system fine-tuned on TACRED data for its ability to extract the same subset of relations present in the NYT dataset, as strong performance on one dataset does not necessarily indicate robust generalisation capabilities. Additionally, techniques such as adversarial attacks (Wu et al., 2019; Gardner et al., 2020; Goel et al., 2021) can further test the true capabilities of RE systems by exposing their vulnerabilities and resilience to challenging scenarios.

5 Conclusion and Future Work

In this work, we have highlighted several limitations in the benchmark documentation and use of traditional leaderboards, particularly those employed for the relation extraction task. Limitations in benchmark documentation include the absence of comprehensive descriptive metadata, such as the source of the data or details regarding inter-annotator agreement, as well as an absence of clear articulation of the dataset’s inherent limitations, such as large class imbalances and potential noise. Furthermore, there is often insufficient discussion on methods to mitigate these issues.

Evaluating systems based on these RE benchmarks inherently necessitates addressing the problems associated with insufficient documentation of the benchmarks. For instance, traditional leaderboards, such as those on PaperswithCode, that play a significant role in advancing NLP, typically rely on holistic metrics like F1-score. However, these metrics fail to capture the complexity of the relation extraction task, especially in scenarios involving a large number of labels and highly imbalanced datasets, such as TACRED, where most instances belong to a ‘no_relation’ class. Additionally, papers reporting new SOTA results on RE benchmarks like NYT and TACRED often focus exclusively on aggregate metrics, neglecting class-based metrics, which obscures the nuanced performance of models across different relation types.

This paper does not intend to undermine the significance and value of existing benchmarks such as TACRED or NYT, which are crucial for the evaluation of models in the field, as well as the development of new SOTA approaches. Instead, given the evolving perspective on data used in training deep learning models, our objective is to propose

avenues for improving the documentation of benchmark creation processes, which would in turn help to better assess the generalisation capabilities of RE models. Additionally, we also aim to motivate the adoption of more rigorous evaluation practices, encouraging researchers to move beyond the limited scope of only reporting metrics such as aggregate F1-score, precision, and recall. This is particularly important in NLP tasks such as relation extraction, where the complexity is exacerbated by the presence of a large number of relations.

It is also crucial to recognise that high performance on a specific RE benchmark, such as TACRED, TACRED-RE, or NYT, reflects only a model’s ability to handle a subset of all possible relations. Furthermore, even if a system performs well on a given subset of relations, it may struggle significantly when extracting the same relations from out-of-distribution data.

Our focus should not solely be on the development of new approaches, but also on critically analysing our systems and recognising the limitations of the data used for their evaluation. This critical perspective is essential for advancing the field and ensuring that our models are robust and generalisable.

Finally, this work serves as a position paper that highlights several issues in the creation of RE benchmarks and the practices surrounding leaderboard evaluations. We acknowledge the limitations of this work, particularly the lack of extensive quantitative evidence. In our future research, we aim to conduct a comprehensive cross-dataset evaluation of RE systems on the benchmarks discussed. Such an evaluation will provide empirical support for the concerns raised and offer a more reliable assessment of the generalisation capabilities of current RE systems.

References

- Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. [TACRED revisited: A thorough evaluation of the TACRED relation extraction task](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. [Freebase: a collaboratively created graph database for structuring human knowledge](#). In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.
- François Chollet. 2019. [On the measure of intelligence](#). *ArXiv*, abs/1911.01547.
- Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. [The benchmark lottery](#). *Preprint*, arXiv:2107.07002.
- Sefika Efeoglu and Adrian Paschke. 2024. [Retrieval-augmented generation-based relation extraction](#). *Preprint*, arXiv:2404.13397.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. [Robustness gym: Unifying the NLP evaluation landscape](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- James Y. Huang, Bangzheng Li, Jiashu Xu, and Muhao Chen. 2022. [Unified semantic typing with meaningful label inference](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Seattle, United States. Association for Computational Linguistics.
- Pere-Lluís Hugué Cabot and Roberto Navigli. 2021. [REBEL: Relation extraction by end-to-end language](#)

- generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, et al. 2023. [A taxonomy and review of generalization research in nlp](#). *Nature Machine Intelligence*, 5(10):1161–1174.
- Mohamad Yaser Jaradeh, Allard Oelen, Manuel Prinz, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: A system walkthrough. In *Digital Libraries for Open Knowledge*, pages 348–351, Cham. Springer International Publishing.
- Venelin Kovatchev and Matthew Lease. 2024. [Benchmark transparency: Measuring the impact of data on evaluation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1536–1551, Mexico City, Mexico. Association for Computational Linguistics.
- Tal Linzen. 2020. [How can we accelerate progress towards human-like linguistic generalization?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.
- Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. [ExplainsBoard: An explainable leaderboard for NLP](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Riccardo Orlando, Pere-Lluís Hugué Cabot, Edoardo Barba, and Roberto Navigli. 2024. [ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14114–14132, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Jaihyun Park and Sullam Jeoung. 2022. [Raison d’être of the benchmark dataset: A survey of current practices of benchmark dataset sharing platforms](#). In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.
- Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. 2021. [Mitigating dataset harms requires stewardship: Lessons from 1000 papers](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Alistair Plum, Tharindu Ranasinghe, Spencer Jones, Constantin Orasan, and Ruslan Mitkov. 2022. [Biographical semi-supervised relation extraction dataset](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 3121–3130, New York, NY, USA. Association for Computing Machinery.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should that change NLP leaderboards?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. [GoLLIE: Annotation guidelines improve zero-shot information-extraction](#). In *The Twelfth International Conference on Learning Representations*.
- Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.
- Prerna Singh. 2023. [Systematic review of data-centric approaches in artificial intelligence and machine learning](#). *Data Science and Management*, 6(3):144–157.
- Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. [UniRel: Unified representation and interaction for joint relational triple extraction](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021. [Zero-shot information extraction as a unified text-to-triple translation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. [DeepStruct: Pre-training of language models for structure prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. [Errudite: Scalable, reproducible, and testable error analysis](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy. Association for Computational Linguistics.
- Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. [Revisiting the negative data of distantly supervised relation extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3572–3581, Online. Association for Computational Linguistics.
- Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. [Noise mitigation for neural entity typing and relation extraction](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194, Valencia, Spain. Association for Computational Linguistics.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A Dataset Statistics: Class Distribution

Table 3: TACRED Dataset

Relation	# of Samples
no_relation	84,491
per:title	3862
org:top_members/employees	2770
per:employee_of	2163
org:alternate_names	1359
per:age	833
per:countries_of_residence	819
org:country_of_headquarters	753
per:cities_of_residence	742
per:origin	667
org:city_of_headquarters	573
per:stateorprovinces_of_residence	484
per:spouse	483
org:subsidiaries	453
org:parents	444
per:date_of_death	394
org:stateorprovince_of_headquarters	350
per:children	347
per:cause_of_death	337
per:other_family	319
per:parents	296
org:members	286
per:charges	280
org:founded_by	268
per:siblings	250
per:schools_attended	229
per:city_of_death	227
org:website	223
org:member_of	171
org:founded	166
per:religion	153
per:alternate_names	153
org:shareholders	144
org:political/religious_affiliation	125
org:number_of_employees/members	121
per:stateorprovince_of_death	104
per:date_of_birth	103
per:city_of_birth	103
per:stateorprovince_of_birth	72
per:country_of_death	61
per:country_of_birth	53
org:dissolved	33
Positive Samples	21,773
Negative Samples	84,491
Total	106,264

Table 4: NYT Dataset

Relation	# of Samples
None	170,021
/location/location/contains	44,490
/location/country/capital	7267
/people/person/nationality	7244
/people/person/place_lived	7015
/location/administrative_division/country	5951
/location/country/administrative_divisions	5851
/business/person/company	5421
/location/neighborhood/neighborhood_of	5082
/people/person/place_of_birth	3133
/people/deceased_person/place_of_death	1914
/business/company/founders	767
/people/person/children	487
/business/company/place_founded	414
/business/company/major_shareholders	282
/business/company_shareholder/major_shareholders	282
/sports/sports_team_location/teams	218
/sports/sports_team/location	218
/people/person/religion	67
/business/company/advisors	45
/people/ethnicity/geographic_distribution	33
/people/ethnicity/people	21
/people/person/ethnicity	21
/people/person/profession	2
/business/company/industry	1
Positive Samples	96,228
Negative Samples	170,021
Total	266,249