# Zero-shot Entailment of Leaderboards for Empirical AI Research

Salomon Kabongo
L3S Research Center,
Leibniz University of Hannover
Germany
kabenamualu@l3s.de

Jennifer D'Souza
TIB Leibniz Information Centre for
Science and Technology
Germany
jennifer.dsouza@tib.eu

Sören Auer
TIB Leibniz Information Centre for
Science and Technology
Germany
auer@tib.eu

## ABSTRACT

We present a large-scale empirical investigation of the zero-shot learning phenomena in a specific recognizing textual entailment (RTE) task category, i.e., the automated mining of LEADERBOARDS for Empirical AI Research. The prior reported state-of-the-art models for LEADERBOARDS extraction formulated as an RTE task in a non-zero-shot setting are promising with above 90% reported performances. However, a central research question remains unexamined: *did the models actually learn entailment?* Thus, for the experiments in this paper, two prior reported state-of-the-art models are tested out-of-the-box for their ability to *generalize* or their capacity for *entailment*, given LEADERBOARD labels that were unseen during training. We hypothesize that if the models learned entailment, their zero-shot performances can be expected to be moderately high as well–perhaps, concretely, better than chance. As a result of this work, a zero-shot labeled dataset is created via distant labeling, formulating the LEADERBOARD extraction RTE task.

## CCS CONCEPTS

• **Computing methodologies → Neural networks**.

## KEYWORDS

Natural-Language-Inference, Entailment, Leaderboard, Information-Extraction

## 1 INTRODUCTION

The recognizing textual entailment (RTE) task is defined as "a directional relation between two text fragments $T - H$, called text ($T$, the entailing text or context or premise), and hypothesis ($H$, the entailed text), so that a human being, with common understanding of language and common background knowledge, can infer that $H$ is most likely true on the basis of the content of $T$" [2–4, 7, 7–9]. Proposed as an official community shared task for the first time in 2005 [7], RTE has enjoyed a constantly growing popularity [6] in the NLP community, as it seems to work as a common framework in which to analyze, compare, and evaluate different techniques used in NLP applications to deal with semantic inference, a common issue shared by many NLP applications. In other words, different NLP applications such as reading comprehension, question answering, information extraction, machine translation, etc., can be recast [10, 18, 22] as an RTE task, thereby placing all tasks within a common evaluation paradigm.

In the realm of the RTE tasks collection, the task of automated mining of LEADERBOARDS for Empirical AI Research currently reports state-of-the-art (SOTA) performances above 90% F1 as an RTE task [14]. The extraction RTE task itself is formulated as follows. For LEADERBOARD RTE, in each $T - H$ instance, the text $T$ corresponds
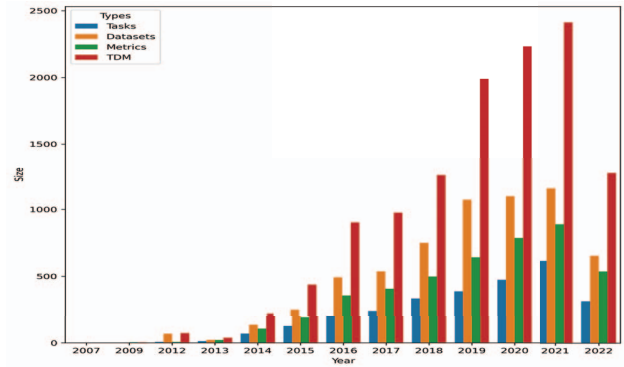


**Figure 1: Rate of introduction of new tasks, datasets, metrics, and (task,dataset,metric) as tuples between 2007 and 2022.**

to the scholarly article or a representation of the scholarly article from which the LEADERBOARD should be extracted. The hypothesis $H$ corresponds to a (task, dataset, metric) tuple, TDM henceforward, obtained from a knowledge base (KB) of TDMs. In the context of the LEADERBOARD extraction task, the extraction objective hypothesis $H$ has been variously formulated in prior work. E.g., either as (task, dataset, metric, method) [13] or (task, dataset, metric, score) [11]. However, since the reported SOTA performance addresses the TDM triple objective, this is the focus of this work.

With respect to the automated mining of LEADERBOARDS from scholarly articles task, there has been a recent surge in interest [11–14, 16, 20]. In this era of the publications deluge where millions of articles are published each year [17], it is no longer feasible for researchers to keep track of innovations by the traditional method of reading dozens of articles. Herein, LEADERBOARDS have proven a smart information access technology as easy progress trackers of models developed, for the empirical AI domain, in particular. LEADERBOARDS are often maintained in websites where models, their results, evaluation metrics, evaluation datasets, and code urls are reported and where specific research problems each have their own LEADERBOARD. With the help of LEADERBOARDS now, researchers interested in developing new models on a problem have easy access to an overview on the current SOTA and performance trends of other models. Traditionally, they would need to scour through the discourse text of several papers and make mental notes of which models reported in which was better–a problem completely alleviated with LEADERBOARDS. Currently, LEADERBOARDS are mainly crowdsourced. E.g., https://nlpprogress.com/, https://paperswithcode.com/, or https://orkg.org/benchmarks. While the community curation of LEADERBOARDS have resulted in thousands of recorded results, note that it does not offer absolute TDM

coverage guarantee. It can be that TDMs that are not noticed will be completely missed. As opposed to crowdsourcing, highly accurate automated extraction methods applied on article collections will not suffer from this limitation. Hence the vision is to complement the community curation of LEADERBOARDS with automated methods.

In lieu of this, a system [14] reporting TDM extraction performances above 90% would make it seem like a practical working solution is found. However, there remains a caveat in the model evaluation. The LEADERBOARD extraction models have not been evaluated on unseen TDM labels. The RTE task, in the prior reported work [14], has only been evaluated on TDM labels that were seen during training. This leads to the formulation of the overarching research question (RQ) of this work: *did the models actually learn entailment?* To obtain concrete insights for the RQ, the strategy we adopt is to test the two SOTA models out-of-the-box, namely ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$, in an exclusive zero-shot TDM extraction RTE setting. Our hypothesis is that if the models actually learnt an entailment task, they will be able to perform at least better than chance on novel unseen TDMs.

Testing the models entailment ability is important for two main reasons. 1) The rapid growth rates of modern science is predicated on millions of scholarly articles published each year [5]. This includes empirical investigations of new models released on different combinations of novel introduced tasks or datasets or metrics. To offer concrete insights into the novelty, Figure 1 illustrates the release rates of tasks, datasets, metrics, respectively, and TDM tuples per year. Thus automated extraction methods need to identify TDM tuples as they are introduced in the community without the need to go through the costly process of repeated retraining of the models as and when new TDM tuples are introduced. And, 2) consequently, one of the main advantages of training an RTE model is its ability to entail new information. Given that the SOTA LEADERBOARD extractor [14] is trained as an RTE model, if it successfully learnt an entailment task, it should be able to extract new TDMs. Thus in the face of the high number of novel TDMs introduced, only the TDMs KB would need to be expanded with new TDMs and the model would not need frequent retraining. To this end, this short paper is then a focused empirical analysis of two state-of-the-art LEADER-BOARD extractors called ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ [14] that are originally trained as RTE models and, in this paper, are exclusively tested for their entailment abilities of new TDMs that were unseen during training. Thus, we investigate ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ in a zero-shot RTE experimental setting.

## 2 RELATED WORK

**Existing LEADERBOARD extraction objectives.** The task is addressed via two different objectives. As 1) information extraction (IE) and 2) entailment (RTE). As an IE task, traditional pipelined based extractors have been implemented. For instance, the SciREX [13] framework, which aims to extract (dataset, metric, task, method) tuples, address it via five IE pipelined modules. They are: mention identification, salient mention identification, pairwise coreference resolution, mention clustering, and document-based 4-ary relation identification. Posited as document-level IE, it uses a 2-level BERT+BiLSTM method to get token representations which are passed to a CRF layer to identify mentions. In the next step, each

mention is classified as being salient or not. Next, a coreference model is trained to cluster these mentions into entities. A final classification layer predicts relationships between 4-tuples of entities represented by their cluster embeddings. A closely related system, CitationIE [20], extends the SciREX architecture by incorporating structural and textual citation information into the mention identification, salient entity classification, and and relation extraction modules. A highlight of these IE models is that the relations need not be provided in advance. Thus when trained on a sizeable proportion of data, these models should become capable of extracting relations between previously unseen entity tuples. But, based on their end-to-end pipelined extraction experimental results, however, these models report less than 25% F1 which is contrary to expectations making the current IE-based models highly unsuitable for use in practice.

On the other hand, works following the RTE objective depend on a predefined KB of TDM tuples as LEADERBOARDS used as candidate hypotheses ($H$) in $T - H$ pairs, instantiated as a paper text ($T$) and a TDM tuple ($H$) to determine one or more TDM hypotheses that can be entailed by a paper text. As the paper text representation $T$, the IBM-TDMS work [11] defined a shorter representation feature instead of simply using the full-text of the article. As $T$ they only use a selection of the paper based on certain parts of the article such as abstract, table captions, experimental section, etc., where LEADERBOARDS or TDMs are most likely to be mentioned. They called this feature the DocTAET representation. Each paper was then classified with TDM tuple(s) or unknown, defined as an entailment between DocTAET ($T$) and the TDM ($H$) tuples. See Table 1 for concrete examples of the data instances. Their model relied on the BERT$_{base}$ transformer encoder architecture [15] with a final classification layer for the entailed or not entailed decision. For a corpus of 332 papers, the TDM KB comprised 77 distinct tuples. They reported 67.8% in F1–a significant improvement over the IE models performing at 25% F1. The following work, called ORKG-TDM [14], built upon the work by IBM-TDMS. They used the same RTE task setup as well as the DocTAET paper text representation. However, they performed the task on a significantly larger dataset of 5,361 papers with a TDM KB of 1,850 unique instances. Further, they contrasted three distinct strategies of transformer models [19]: BERT$_{base}$ [15], scientific BERT [1], and XLNet [24]. To date they report SOTA performance on the task with above 90% in F1.

Given the promise shown by RTE models over IE, by performing the SOTA RTE zero-shot evaluation in this work, we provide concrete evidence in the community on the robustness of the LEADER-BOARD extraction RTE models as actual entailment models.

**Diverse NLP tasks recast as recognizing textual entailment (RTE).** Entailment is seen as a core function implicit to various NLP applications to successfully perform a task. It is possible for various semantic classes of NLP tasks to be explicitly recast as an entailment task thus offering a common framework to analyze, compare, and evaluate different techniques. Since 2005, the task has been performed in annual challenges, called the Recognizing Textual Entailment (RTE) challenges, which have helped foster the interest of the research community in textual entailment. In these RTE-1 to 7 challenges [2–4, 7, 7–9], common NLP applications like information retrieval (IR), comparable documents (CD), reading

| Text ($T$) | Hypothesis ($H$) |
|---|---|
| 3D Instances as 1D Kernels ? We introduce a 3D instance representation … clustering algorithms in standard 3D instance segmentation pipelines. Results show that DKNet outperforms the state of the arts on both ScanNetV2 and S3DIS datasets .. The others are evaluated via 6- fold cross validation. mRec mPre mWCov mCov Table 5. | 3D Instance Segmentation; S3DIS; mCov |
| Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors Oriented object detection in aerial images is a challenging task … We evaluate our method on two public aerial image datasets: DOTA and HRSC2016 We use DOTA-v1.0 dataset … second best detection results in each column. 2 BC ST BD SV Plane Harbor LV TC RA SBF GTF Ship HC mAP SP Bridge 2 FPS Image Size AP Table 3 | Object Detection In Aerial Images; DOTA; mAP |

**Table 1: Recognizing textual entailment (RTE) instances as text-hypothesis $T − H$ pairs for the task of automated mining of LEADERBOARDS. The Table shows one LEADERBOARD annotation for the two papers "3D Instances as 1D Kernels" [23] and "Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors" [25], respectively. LEADERBOARD constitute the hypothesis ($H$) in the RTE task and are expressed as ($task$; $dataset$; $metric$) tuples. On the other hand, the text ($T$) shown in the first column are snippets of the DocTAET compressed paper representation comprising only the extracted paper title, abstract, experimental setup, and tabular information instead of the paper's full text.**

comprehension (RC), question answering (QA), information extraction (IE), machine translation (MT), and summarization (SUM) were recast as entailment tasks. The $T − H$ pairs were created by expert annotators in the different NLP application settings. E.g., considering an entailment instance for the QA task, $T$ was the text where the answer is found and $H$ is the question reformulated as an assertion with the answer in it. As another example, in the SUM task context, $T$ is a text description whose summary is the assertion hypothesis $H$. The RTE task itself then required the participating systems to decide, given two text snippets $T$ and $H$, whether $T$ entails $H$, where the $T − H$ pairs were recast equivalent and explicit entailment formulations of the different popular NLP applications. Three main aspects that informed the evolution of the tasks were: 1) RTE first postulated as two-way classification (entailment and unknown) was extended as a three-way classification task (entailment, contradiction and unknown). Further, 2) to situate RTE realistically, the length of the Ts were made longer in later task runs. In RTE-5, $T$s had lengths up to 100 words, whereas in RTE-4 the average length was about 40 words. This length was meant to represent the average portion of the source document that a reader would naturally select, such as a paragraph or a group of related sentences. Also, 3) later variants of the RTE task focused on search applications in the context of summarization, to retrieve all relevant texts $T$ from a collection of documents given hypothesis $H$ as queries. The development of search technologies in the later tasks further bolstered the interest of the research communities to investigating the recast RTE task formulations.

Inspired from the RTE challenges, there were other initiatives that recast common NLP tasks as RTE. To this end, the Diverse Natural Language Inference Collection (DNC) [18, 22] recast seven semantic phenomena from a total of 13 datasets into labeled RTE examples. The seven recast semantic phenomenon were the following: event factuality (EF), named entity recognition (NER), gendered anaphora resolution (GAR), lexicosyntactic inference (Lex), figurative language (Puns), relation extraction (RE), and subjectivity (Sentiment). Each of these seven tasks were reformulated in terms of the RTE $T − H$ tuples, where $T$ was the task sentence itself and the hypothesis $H$, informed by the underlying task, was a short,

manually constructed sentence having an entailed or not-entailed resolution. E.g., for Lex, the premise is a sentence with various kinds of verb constructions about an event. Then each context sentence was paired with three hypotheses: a. That thing happened, b. That thing may or may not have happened, and c. That thing didn't happen. And the entailment task was entailed or not for each of the $T − H$ a, b, and c pairs.

Inspired by these various RTE task formulations, the LEADERBOARD RTE task defined the hypothesis $H$ as the TDM tuple to be entailed from the text representation $T$ posited as the shorter and selective DocTAET feature of the original paper full text. Here, aligned with the later versions of the RTE challenges (RTE-5 onward), we are also testing the models in the realistic setting of longer $T$ from which to entail information. Offering a paragraph as $T$ as opposed to a single sentence has been shown to increase the RTE task complexity.

## 3 ZERO-SHOT EXPERIMENTAL SETUP

### 3.1 Our Zero-shot Corpus of LEADERBOARDS

To test the reference trained models in a zero-shot setting, we first needed to create a suitable experimental corpus. *The essential criteria to be satisfied by the new corpus was that none of the TDMs seen by the models of the reference work [14] during training could be included in the new experimental corpus.* In other words, the experimental corpus for this paper's zero-shot labeling experiments needed to contain only unseen TDM labels for the ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ trained models from the reference work [14]. We addressed this by first downloading a new dump of community-curated papers of LEADERBOARDS (timestamp: Nov. 04, 2022) from https://paperswithcode.com/. Note we rely on the same data source, viz. https://paperswithcode.com/, as that of the reference work [14]. The dump download of the reference work to train the models had a timestamp of May 10, 2021. By a simple difference of the latest dump with the earlier, we first obtained a corpus of papers that were newly introduced and that were not part of the reference work dataset. While this ensured unseen $T$s in the $T − H$ pairs, we also

|  | Macro P | Macro R | Macro F1 | Micro P | Micro R | Micro F1 |
|---|---|---|---|---|---|---|
| ORKG-TDM$_{Bert}$ | | | | | | |
| Fold-1 | 20.1 | 83.4 | 28.9 | 14.1 | 72.9 | 23.6 |
| Fold-2 | 16.2 | 89 | 24.4 | 10.4 | 81.7 | 18.4 |
| Average Fold 1 and Fold 2 | 18.2 | 86.2 | 26.7 | 12.3 | 77.3 | 21.0 |
| ORKG-TDM$_{XLNet}$ | | | | | | |
| Fold-1 | 14.3 | 86.6 | 21.9 | 9.2 | 78.1 | 16.5 |
| Fold-2 | 14.9 | 86.4 | 22.7 | 10.1 | 76.8 | 17.8 |
| Average Fold 1 and Fold 2 | 14.6 | 86.5 | 22.3 | 9.7 | 77.5 | 17.2 |

**Table 2: Zero-shot LEADERBOARD RTE results for the two best models, viz. ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$, from Kabongo et al. 14 with reported performances above 90% in F1 in non-zero-shot experimental evaluations.**

wanted to establish a corpus of unseen $H$s or TDMs to satisfy the zero-shot labeling criteria i.e. the entailment of new TDMs.

The next natural question is *how did we define a zero-shot TDM label?* To arrive at this definition, we made the following critical observation. Had we defined a zero-shot TDM label as one where each of the task, dataset, and metric that constituted a TDM were all unseen in the training dataset of the reference work, this would have resulted in an empty zero-shot TDM labels set. It is a common phenomenon that the ratio of new metrics released every year is almost negligible. Further, the introduction of new datasets for existing tasks is far more frequent than the introduction of new tasks itself. This is reflected in the trends depicted in Figure 1. Thus instead of adopting a strict criteria where all elements of TDM tuple were new, we instead defined a zero-shot TDM tuple as one where any of the task, dataset, or metric were unseen before. The new set of papers were further filtered to satisfy this criteria in their TDM annotations. To better clarify, a paper had to have only zero-shot TDM labels per our adopted definition to qualify as a candidate for the zero-shot corpus of LEADERBOARDS. While this resulted in a corpus of over 3000 papers, we randomly selected a subsample of 1000 papers which had a resulting set of 1925 zero-shot TDM labels including the label "unknown."

Note that this work is relegated purely to test evaluations, and therefore there is no training set. In other words, the sample of 1000 papers with 1925 zero-shot TDM labels constitute a testing corpus on which the ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ trained models from the reference work [14] are tested out-of-the-box.

### 3.2 Two Test LEADERBOARD RTE Models

The two selected test models, viz. ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ uses the standard transformer models [19] fine-tuned for sequence pair classification, used here for RTE, with a [SEP] token between the the DocTAET paper representation as text $T$ and the TDM tuple as hypothesis $H$. The maximum input length is 512 for BERT [15] and 2000 for XLNET [24].

## 4 RESULTS AND DISCUSSION

Experimental results from the ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ applied in a zero-shot RTE setting are shown in Table 2. Since the original models were evaluated in a two-fold setting, in this work, we apply the models from their respective folds on our zero-shot corpus of LEADERBOARDS. This is how two-fold results are reported.

Note the underlying test corpus in each fold is the same, only the trained models per fold differ. Our initial RQ was: *did the models actually learn entailment?* We observe results contrary to our positive expectations based on the models strong performances reported in their non-zero-shot results. ORKG-TDM$_{Bert}$ in a non-zero-shot setting reported 90.8% macro F1 and 91.8% micro F1 [14]. The same model in a zero-shot setting reports performances significantly less than chance at 26.7% macro F1 and 21% micro F1. Likewise, the other SOTA model ORKG-TDM$_{XLNet}$ in a non-zero-shot setting reported 91.2% macro F1 and 92.4% micro F1. This model in the zero-shot setting reported 22.3% macro F1 and 17.2% micro F1. By observing the results in Table 2, we see that the models have a very high recall at the cost of precision. Thus, for each $T$, the models entail most of the TDM hypothesis as true without being able to identify which TDM hypothesis the given $T$ actually entails.

We relegate the following two main reasons for the model low performances. 1) The RTE task is essentially a sequence-pair classification machine learning objective. In other words, whether the model learns a classification task or a true entailment task depends on the nature of the problem and cannot be enforced upon the model. Given the ORKG-TDM$_{Bert}$ and ORKG-TDM$_{XLNet}$ strong performances in the non-zero-shot setting versus poor performances in the zero-shot setting, we hypothesize that the reference models originally learned as a multi-class classification task and not an entailment task in its true sense. 2) The longer DocTAET $T$ representation further poses a challenge for the model to learn an entailment task. We hypothesize that a shorter one or two sentence representation would be a better approach toward realizing a true RTE model for LEADERBOARD extraction.

## 5 CONCLUSION

We have offered a comprehensive and realistic look at LEADERBOARD extraction as an RTE task. We hope the results of this work can offer more informed decisions for future research on this theme. E.g., investigating transformer-model powered prompt-based methods instead that have shown promise as zero-shot learners [21].

# REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3615–3620.

[2] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *Text Analysis Conference*.

[3] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. *Theory and Applications of Categories* (2011).

[4] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. [n. d.]. The Fifth PASCAL Recognizing Textual Entailment Challenge. ([n. d.]).

[5] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 66, 11 (2015), 2215–2222.

[6] Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of Natural Language Inference Data Collection: The way forward?. In *Proceedings of the Computing Natural Language Inference Workshop*.

[7] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment* (Southampton, UK) *(MLCW'05)*. Springer-Verlag, Berlin, Heidelberg, 177–190. https://doi.org/10.1007/11736790_9

[8] Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The Fourth PASCAL Recognizing Textual Entailment Challenge.. In *TAC*.

[9] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, Prague, 1–9. https://aclanthology.org/W07-1401

[10] Oren Glickman. 2006. *Applied textual entailment*. Citeseer.

[11] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5203–5213. https://doi.org/10.18653/v1/P19-1513

[12] Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 707–714.

[13] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7506–7516.

[14] Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021. Automated mining of leaderboards for empirical ai research. In *International Conference on Asian Digital Libraries*. Springer, 453–470.

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. 4171–4186.

[16] Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1885–1895.

[17] Andrew Plume. 2011. Drowning in the publication deluge? *Research Trends* 1, 22 (2011), 7.

[18] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 67–81. https://doi.org/10.18653/v1/D18-1007

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[20] Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the Citation Graph for Scientific Information 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)*. Association for Computational Linguistics, Online, 719–731. https://doi.org/10.18653/v1/2021.acl-long.59

[21] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. [n. d.]. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

[22] Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is Everything: Recasting Semantic Resources into a Unified Evaluation Framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 996–1005. https://aclanthology.org/I17-1100

[23] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022. 3D Instances as 1D Kernels. In *European Conference on Computer Vision*. Springer, 235–252.

[24] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[25] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. 2021. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2150–2159.