Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications

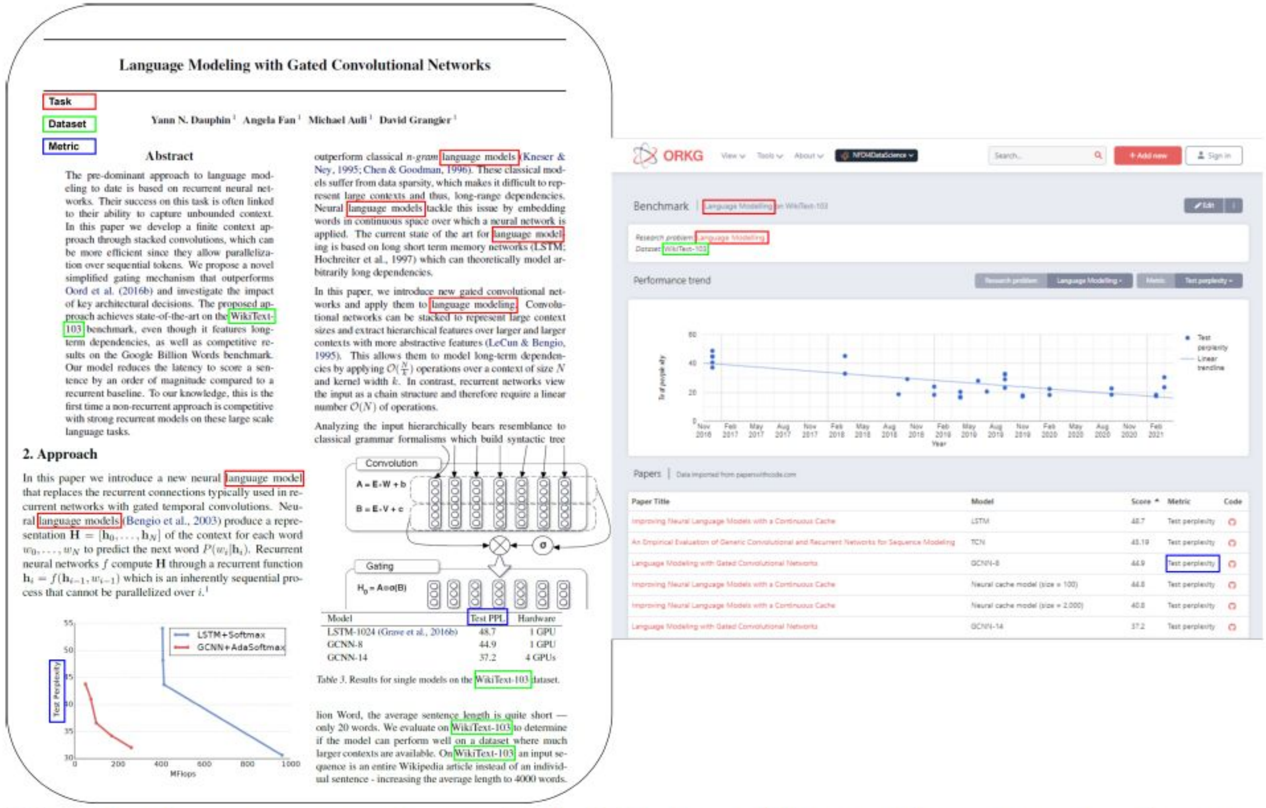
Presented at: CLEF 2024 Conference, SimpleText Track, SimpleText Session 1
Date of event: 9th September 2024
Presented by: Jennifer D’Souza, Junior AI Research Group Leader
Contact: https://www.linkedin.com/in/jennifer-l-dsouza/
Reference: D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

- **Task Definition:**
  - Objective: Extract Task, Dataset, Metric, Score tuples from research papers to automatically construct leaderboards of AI models.
  - To fulfill the objective, systems had to perform the following two tasks:
    1. **classification** – given the full text of an AI scientific paper, classify whether the paper indeed reports model scores on benchmark datasets, and if so,
    2. **information extraction** – extract all pertinent (Task, Dataset, Metric, Score) tuples from the content of the scientific paper to automatically populate leaderboards used to keep track on the latest and greatest AI models.

Background

- As a novel addition to the CLEF 2024 SimpleText Track, “SOTA?” explored the structured scientific information model, as advocated by the Open Research Knowledge Graph (ORKG) project, offering a new perspective on the objective of simplifying scientific information. Specifically, “SOTA?” focused on leaderboards or scoreboards in Artificial Intelligence (AI) research. These leaderboards report new *AI models* and their *scores* in terms of the addressed *tasks*, evaluated *datasets*, and applied evaluation *metrics*.



Background: Enhancing Machine-Actionability of Scientific Knowledge

- **Objective**: Generate structured summaries of scientific texts to improve machine-actionability as an alternative to simplify access to scientific advancements.
  - **Benefits**: Helps in managing the vast number of publications and aids in keeping up with scientific advancements using advanced IT tools.
- **SOTA? as an Exemplary Research-problem-specific Use-case:** AI research leaderboards which track and compare model performances on specific tasks and datasets, providing a structured way to assess advancements in AI. This critical information is often deeply embedded in scholarly AI articles.
  - Thus SimpleText in 2024 introduced “Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications” that handled the automatic text mining of the (Task, Dataset, Metric, Score) tuples from AI articles to automatically build leaderboards, where the leaderboards in turn help researchers to directly stay on track of Ai advancements.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Brief History

- Leaderboards have been traditionally curated by the community. Some examples are:
  - http://nlpprogress.com/
  - https://www.eff.org/ai/metrics
  - Dataset-specific leaderboards https://rajpurkar.github.io/SQuAD-explorer/
  - https://paperswithcode.com/
  - https://orkg.org/benchmarks
- Community curation methods often have some limitations:
  - Coverage: there is no guarantee that all models reported in the scientific literature are reported
  - Standardization: different users might have their own terminology to record the information in the leaderboards. For example, some user might represent a score as a percentage, another user might represent it in decimal format. Thus there is no guarantee that the information recorded in the Leaderboard actually aligns with how the information was reported in the paper.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

- Utilizing text mining techniques allows for a transition from the conventional community-based leaderboard curation to an automated text mining approach. Consequently, the goal of Task 4: SOTA? is to develop systems that can classify whether a scholarly article provided as input to the model reports a (T, D, M, S) or not. And for articles reporting (T, D, M, S), extract all the relevant ones from the paper text.
- **Formalism.**
    - The Task 4: SOTA? task formalism is defined as follows: given the text of a scientific paper A, the goal is to extract its Leaderboards L, where L = {l1, ..., lx} and A can have between one to an undefined number of Leaderboards. Each Leaderboard l comprises the (T, D, M, S) quadruple.
    - Systems were evaluated in two separate evaluation phases:
        - **Evaluation Phase I.** Few-shot (T,D,M,S) extraction: Systems are expected to identify whether an incoming AI paper reports leaderboards or not; and for paper's reporting leaderboards, extract all the pertinent (T, D, M, S) quadruples. The "few-shot" aspect of this subtask is that it involves (T, D, M) labels previously seen in the training dataset.
        - **Evaluation Phase II.** Zero-shot (T,D,M,S) extraction: This is similar to Phase I, but involves a new test dataset containing (T, D, M) tuples that were not seen in the training set, testing the system's ability to handle zero-shot scenarios.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset

- Overall Dataset
  - Papers with Leaderboard Annotations:
    - The corpus included over 8,000 articles, with 7,987 used for training and 994 for testing, divided into 751 for the few-shot setting and 241 for the zero-shot setting.
    - Data Sources
      - Leaderboard annotations from PapersWithCode. Specifically the PwC data downloaded on December 09, 2023 [1]
      - The full-text of the articles was sourced from the arXiv preprint server under CC-BY licenses
        - Each article in the dataset is available in TEI XML format, complete with one or more (T, D, M, S) annotations from PwC
  - Papers without Leaderboards i.e. the “unanswerable” set of papers.
    - Included a set of approximately 4,401 and 648 articles that do not report leaderboards into the train and test sets.
    - These articles were randomly selected by leveraging the arxiv category feature, then filtering it to papers belonging to domains unrelated to AI/ML/Stats. These articles were annotated with the unanswerable label.
- Thus given the overall dataset, systems could perform the expected task i.e. classification and information extraction.
- Dataset release: https://github.com/jd-coderepos/sota

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

- Train Dataset: 7,936 papers annotated with leaderboards, and 4,352 as "unanswerable”.
- Validation Dataset: 51 papers with leaderboard annotations and 49 as "unanswerable”.
- Few-shot test dataset for evaluation phase 1: 753 with leaderboards and 648 as "unanswerable”.
- Zero-shot test dataset for evaluation phase 2: 241 with leaderboards and 548 as "unanswerable”.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 1
SimpleText Task 4: SOTA? dataset statistics displaying unique labels for annotated (Task, Dataset, Metric) elements.

Parameter	Train	Few-shot Test	Zero-shot Test
Unique Tasks	1,372	320	236
Unique Datasets	4,795	935	646
Unique Metrics	2,782	637	397
Unique (Task, Dataset, Metric) triples	11,977	1,900	1,262
Avg. (Task, Dataset, Metric) triples per paper	6.93	5.69	7.53

Table shows the unique mentions of Tasks, Datasets, Metrics across the datasets

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 1
SimpleText Task 4: SOTA? dataset statistics displaying unique labels for annotated (Task, Dataset, Metric) elements.

Parameter	Train	Few-shot Test	Zero-shot Test
Unique Tasks	1,372	320	236
Unique Datasets	4,795	935	646
Unique Metrics	2,782	637	397
Unique (Task, Dataset, Metric) triples	11,977	1,900	1,262
Avg. (Task, Dataset, Metric) triples per paper	6.93	5.69	7.53

- Table shows the unique mentions of Tasks, Datasets, Metrics across the datasets.
- Most pronounced for Datasets, then Metrics, and then Tasks.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 1
SimpleText Task 4: SOTA? dataset statistics displaying unique labels for annotated (Task, Dataset, Metric) elements.

Parameter	Train	Few-shot Test	Zero-shot Test
Unique Tasks	1,372	320	236
Unique Datasets	4,795	935	646
Unique Metrics	2,782	637	397
Unique (Task, Dataset, Metric) triples	11,977	1,900	1,262
Avg. (Task, Dataset, Metric) triples per paper	6.93	5.69	7.53

- Table shows the unique mentions of Tasks, Datasets, Metrics across the datasets.
- Most pronounced for Datasets, then Metrics, and then Tasks.
- This novelty partially stems from the community-curated annotations in the PwC, which result in unnormalized labels. For instance, the metric “F1-score” might be recorded as “F1,” “F-score,” or “F-measure,” and each variation is considered a unique Metric label. This diversity aims to mirror the variability seen in scientific papers, where, to our knowledge, there is no standardized naming convention for these entities.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 2
Ten most common Tasks, Datasets, and Metrics in the SimpleText Task 4: SOTA? training dataset.

Task	**Frequency**	*Dataset*	**Frequency**	*Metric*	**Frequency**
Image Classification	2,273	ImageNet	1,603	Accuracy	4,383
Atari Games	1,448	COCO Test-Dev	792	Score	1,515
Node Classification	1,113	Human3.6M	624	F1	1,384
Object Detection	1,001	CIFAR-10	585	PSNR	1,144
Video Retrieval	997	COCO Minival	310	MAP	1,068
Link Prediction	941	YouTube-VOS 2018	295	MIoU	862
Semantic Segmentation	901	CIFAR-100	252	SSIM	799
Semi-supervised Video Object Segmentation	890	MSR-VTT-1kA	247	Top 1 Accuracy	789
3D Human Pose Estimation	889	FB15k-237	244	1:1 Accuracy	787
Question Answering	866	MSU Super-Resolution for Video Compression	225	Number of Params	759

Table 3
Ten most common (Task, Dataset, Metric) triples in the SimpleText Task 4: SOTA? training dataset.

(*Task*, *Dataset*, *Metric*)	**Frequency**
(Image classification, ImageNet, Top 1 accuracy)	524
(Image classification, ImageNet, Number of params)	313
(Image classification, ImageNet, GFLOPs)	256
(3D human pose estimation, Human3.6M, Average MPJPE)	197
(Image classification, CIFAR-10, Percentage correct)	128
(Action classification, Kinetics-400, ACC@1)	108
(Object detection, COCO test-dev, Box mAP)	106
(Image classification, CIFAR-100, Percentage correct)	105
(Semantic segmentation, ADE20K, Validation mIoU)	92
(Neural architecture search, ImageNet, Top-1 error)	83

- Tables 2 and 3 display the top 10 most frequent (Task, Dataset, Metric) annotations in the SOTA? dataset, both as individual elements and as combined triples.
- This may also indicate a prevailing research trend within the scientific community: “Image Classification” is a commonly addressed task, and the “ImageNet” dataset is frequently used to develop or evaluate systems, often employing variants of the “accuracy” metric.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 4
SimpleText Task 4: SOTA? dataset statistics showing the proportion of annotated elements (Task, Dataset, Metric, Score), where the annotation label text exactly matches the text found within the paper.

Dataset Count Parameter	Train	Few-shot Test	Zero-shot Test
Unique *Tasks* per Paper	10,810	1,008	351
Unique found-in-paper *Tasks* per Paper	6,512	649	222
Ratio *Tasks*	0.6024	0.6438	0.6325
Unique *Datasets* per Paper	21,278	1,937	777
Unique found-in-paper *Datasets* per Paper	9,677	816	328
Ratio *Datasets*	0.4548	0.4213	0.4221
Unique *Metrics* per Paper	23,220	2,136	702
Unique found-in-paper *Metrics* per Paper	9,913	861	340
Ratio *Metrics*	0.4269	0.4031	0.4843
Unique *Scores* per Paper	52,092	4,110	1,688
Unique found-in-paper *Scores* per Paper	30,660	2,266	911
Ratio *Scores*	0.5886	0.5513	0.5462

- Table 4 offers insights to what extent of the annotated leaderboards, the respective (T, D, M, S) labels were found in the underlying source text across the Train and the two Test datasets.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 4
SimpleText Task 4: SOTA? dataset statistics showing the proportion of annotated elements (Task, Dataset, Metric, Score), where the annotation label text exactly matches the text found within the paper.

Dataset Count Parameter	Train	Few-shot Test	Zero-shot Test
Unique *Tasks* per Paper	10,810	1,008	351
Unique found-in-paper *Tasks* per Paper	6,512	649	222
Ratio *Tasks*	0.6024	0.6438	0.6325
Unique *Datasets* per Paper	21,278	1,937	777
Unique found-in-paper *Datasets* per Paper	9,677	816	328
Ratio *Datasets*	0.4548	0.4213	0.4221
Unique *Metrics* per Paper	23,220	2,136	702
Unique found-in-paper *Metrics* per Paper	9,913	861	340
Ratio *Metrics*	0.4269	0.4031	0.4843
Unique *Scores* per Paper	52,092	4,110	1,688
Unique found-in-paper *Scores* per Paper	30,660	2,266	911
Ratio *Scores*	0.5886	0.5513	0.5462

- Table 4 offers insights to what extent of the annotated leaderboards, the respective (T, D, M, S) labels were found in the underlying source text across the Train and the two Test datasets.
- In the training dataset, we see: 60.24% for Tasks, 58.86% for Scores, 45.48% for Datasets, and 42.69% for Metrics. This data indicates that Metrics exhibit the greatest inconsistency in annotation labels, followed by Datasets, Scores, and Tasks.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 4
SimpleText Task 4: SOTA? dataset statistics showing the proportion of annotated elements (Task, Dataset, Metric, Score), where the annotation label text exactly matches the text found within the paper.

Dataset Count Parameter	Train	Few-shot Test	Zero-shot Test
Unique *Tasks* per Paper	10,810	1,008	351
Unique found-in-paper *Tasks* per Paper	6,512	649	222
Ratio *Tasks*	0.6024	0.6438	0.6325
Unique *Datasets* per Paper	21,278	1,937	777
Unique found-in-paper *Datasets* per Paper	9,677	816	328
Ratio *Datasets*	0.4548	0.4213	0.4221
Unique *Metrics* per Paper	23,220	2,136	702
Unique found-in-paper *Metrics* per Paper	9,913	861	340
Ratio *Metrics*	0.4269	0.4031	0.4843
Unique *Scores* per Paper	52,092	4,110	1,688
Unique found-in-paper *Scores* per Paper	30,660	2,266	911
Ratio *Scores*	0.5886	0.5513	0.5462

- Table 4 offers insights to what extent of the annotated leaderboards, the respective (T, D, M, S) labels were found in the underlying source text across the Train and the two Test datasets.
- In the training dataset, we see: 60.24% for Tasks, 58.86% for Scores, 45.48% for Datasets, and 42.69% for Metrics. This data indicates that Metrics exhibit the greatest inconsistency in annotation labels, followed by Datasets, Scores, and Tasks.
- This is a crucial perspective in interpreting the performance of participant systems in this year's Task 4: SOTA? dataset which presents the most variability in annotations in the training and evaluation of participant systems which in turn can account for lower reported scores.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Dataset Statistics

Table 4
SimpleText Task 4: SOTA? dataset statistics showing the proportion of annotated elements (Task, Dataset, Metric, Score), where the annotation label text exactly matches the text found within the paper.

Dataset Count Parameter	Train	Few-shot Test	Zero-shot Test
Unique *Tasks* per Paper	10,810	1,008	351
Unique found-in-paper *Tasks* per Paper	6,512	649	222
Ratio *Tasks*	0.6024	0.6438	0.6325
Unique *Datasets* per Paper	21,278	1,937	777
Unique found-in-paper *Datasets* per Paper	9,677	816	328
Ratio *Datasets*	0.4548	0.4213	0.4221
Unique *Metrics* per Paper	23,220	2,136	702
Unique found-in-paper *Metrics* per Paper	9,913	861	340
Ratio *Metrics*	0.4269	0.4031	0.4843
Unique *Scores* per Paper	52,092	4,110	1,688
Unique found-in-paper *Scores* per Paper	30,660	2,266	911
Ratio *Scores*	0.5886	0.5513	0.5462

- Table 4 offers insights to what extent of the annotated leaderboards, the respective (T, D, M, S) labels were found in the underlying source text across the Train and the two Test datasets.
- In the training dataset, we see: 60.24% for Tasks, 58.86% for Scores, 45.48% for Datasets, and 42.69% for Metrics. This data indicates that Metrics exhibit the greatest inconsistency in annotation labels, followed by Datasets, Scores, and Tasks.
- This is a crucial perspective in interpreting the performance of participant systems in this year’s Task 4: SOTA? dataset which presents the most variability in annotations in the training and evaluation of participant systems which in turn can account for lower reported scores.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Submission Format


```
[{'LEADERBOARD':
  {'Task': 'Semi-Supervised Video Object Segmentation',
   'Metric': 'Jaccard (Mean)',
   'Score': '71.6'
   }
   },
   {'LEADERBOARD':
    {'Task': 'Semi-Supervised Video Object Segmentation',
     'Dataset': 'DAVIS 2017 (val)',
     'Metric': 'F-measure (Mean)',
     'Score': '77.7'
     }
    },
    {'LEADERBOARD':
     {'Task': 'Semi-Supervised Video Object Segmentation',
      'Dataset': 'DAVIS 2017 (val)',
      'Metric': 'J&F',
      'Score': '74.65'
      }
    },
    {'LEADERBOARD':
     {'Task': 'Visual Object Tracking',
      'Dataset': 'YouTube-VOS 2018',
      'Metric': 'Jaccard (Seen)',
      'Score': '73.5'
      }
    },
    {'LEADERBOARD':
     {'Task': 'Visual Object Tracking',
      'Dataset': 'YouTube-VOS 2018',
      'Metric': 'Jaccard (Unseen)',
      'Score': '64.3'
      }
    }
]
```


Figure 1: Submission format example for one paper containing (T, D, M, S) annotations. This file is publicly released online and shows the leaderboard annotations for the paper titled "Proposal, tracking and segmentation (pts): A cascaded network for video object segmentation" [13].

- In the evaluation phases, participants were expected to produce annotation files for each paper according to a prescribed JSON format (shown in the image).

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Evaluation Metrics

- There were 3 main categories of evaluations:
  a. **Classification Accuracy:** This metric measured the accuracy with which the participant systems identified the "unanswerable" papers i.e. papers without leaderboards compared with the gold-standard.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Evaluation Metrics

- There were 3 main categories of evaluations:
  a. **Classification Accuracy:** This metric measured the accuracy with which the participant systems identified the “unanswerable” papers i.e. papers without leaderboards compared with the gold-standard.
  b. **Summarization Rouge:** These metrics provide a quantitative assessment of the similarity between the generated and reference summaries, helping researchers and developers evaluate and compare the effectiveness of different summarization approaches. Analogously, we treated the (T, D, M, S) extraction task as analogous to a summarization objective and hence reported system overall extraction performance based on various ROUGE summarization metrics.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Evaluation Metrics

- There were 3 main categories of evaluations:
  a. **Classification Accuracy:** This metric measured the accuracy with which the participant systems identified the "unanswerable" papers i.e. papers without leaderboards compared with the gold-standard.
  b. **Summarization Rouge:** These metrics provide a quantitative assessment of the similarity between the generated and reference summaries, helping researchers and developers evaluate and compare the effectiveness of different summarization approaches. Analogously, we treated the (T, D, M, S) extraction task as analogous to a summarization objective and hence reported system overall extraction performance based on various ROUGE summarization metrics.
  c. **Per (T, D, M, S) Element-wise Extraction F1-score:** In this evaluation category, we evaluated the model JSON output in a fine-grained manner w.r.t. each of the individual (T, D, M, S) elements and overall for which we reported the results in terms of the standard recall, precision, and F1 score. In addition, we reported exact match and partial match scores.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Task 4 Evaluation Metrics

- There were 3 main categories of evaluations:
    a. **Classification Accuracy:** This metric measured the accuracy with which the participant systems identified the “unanswerable” papers i.e. papers without leaderboards compared with the gold-standard.
    b. **Summarization Rouge:** These metrics provide a quantitative assessment of the similarity between the generated and reference summaries, helping researchers and developers evaluate and compare the effectiveness of different summarization approaches. Analogously, we treated the (T, D, M, S) extraction task as analogous to a summarization objective and hence reported system overall extraction performance based on various ROUGE summarization metrics.
    c. **Per (T, D, M, S) Element-wise Extraction F1-score:** In this evaluation category, we evaluated the model JSON output in a fine-grained manner w.r.t. each of the individual (T, D, M, S) elements and overall for which we reported the results in terms of the standard recall, precision, and F1 score. In addition, we reported exact match and partial match scores.
        - The script operated in two steps: it first compared each predicted (T, D, M, S) unit to the gold standard to find the best match, and then it calculated the individual element-wise extraction measures to determine the overall system recall, precision, and F1-score.
        - Evaluation script is publicly released https://github.com/Kabongosalomon/scoring_program/blob/main/evaluation.py

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- 2 participant teams submitted 36 runs in total.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 1. Team AMATU (Staudinger et al., 2024) | Technical University of Vienna, Austria (TU Wien)
  a. Submitted a total of three runs for the few-shot evaluation phase 1 and nine runs for the zero-shot evaluation phase 2.

References

Staudinger, M., El-Ebshihy, A., Ningtyas, A. M., Piroi, F., & Hanbury, A. (2024). AMATU@Simpletext2024: Are LLMs alone any good for scientific entity extraction? In G. Faggioli, N. Ferro, P. Galuščáková, & A. G. S. de Herrera (Eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 1. Team AMATU (Staudinger et al., 2024) | Technical University of Vienna, Austria (TU Wien)
    a. Submitted a total of three runs for the few-shot evaluation phase 1 and nine runs for the zero-shot evaluation phase 2.
    b. Approach: Two main categories:

References
Staudinger, M., El-Ebshihy, A., Ningtyas, A. M., Piroi, F., & Hanbury, A. (2024). AMATU@Simpletext2024: Are LLMs alone any good for scientific entity extraction? In G. Faggioli, N. Ferro, P. Galuščáková, & A. G. S. de Herrera (Eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 1. Team AMATU (Staudinger et al., 2024) | Technical University of Vienna, Austria (TU Wien)
    a. Submitted a total of three runs for the few-shot evaluation phase 1 and nine runs for the zero-shot evaluation phase 2.
    b. Approach: Two main categories:
        i. A pure pattern-based approach inspired after AxCell (Kardas et al, 2020), and

References

Staudinger, M., El-Ebshihy, A., Ningtyas, A. M., Piroi, F., & Hanbury, A. (2024). AMATU@Simpletext2024: Are LLMs alone any good for scientific entity extraction? In G. Faggioli, N. Ferro, P. Galuščáková, & A. G. S. de Herrera (Eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

M. Kardas, P. Czapla, P. Stenetorp, S. Ruder, S. Riedel, R. Taylor, R. Stojnic, Axcell: Automatic extraction of results from machine learning papers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8580–8594.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 1. Team AMATU (Staudinger et al., 2024) | Technical University of Vienna, Austria (TU Wien)
  a. Submitted a total of three runs for the few-shot evaluation phase 1 and nine runs for the zero-shot evaluation phase 2.
  b. Approach: Two main categories:
     i. A pure pattern-based approach inspired after AxCell (Kardas et al, 2020), and
     ii. An AI-based approach using LLMs with a zero-shot prompt and a few-shot prompt tested for GPT-3.5 and Mistral-7B out-of-the-box. Here, they also experimented with variants on the input scholarly article text from which the (T, D, M, S) annotations were expected to be extracted. This we generally refer to as the context. Two context variants were tried: 1) full paper text and 2) only the text from sections referring to experiments and results, in addition to the abstract, which was pre-extracted inspired by the Argumentative Zoning (AZ) method (Teufel et al., 1999).

References

Staudinger, M., El-Ebshihy, A., Ningtyas, A. M., Piroi, F., & Hanbury, A. (2024). AMATU@Simpletext2024: Are LLMs alone any good for scientific entity extraction? In G. Faggioli, N. Ferro, P. Galuščáková, & A. G. S. de Herrera (Eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

M. Kardas, P. Czapla, P. Stenetorp, S. Ruder, S. Riedel, R. Taylor, R. Stojnic, Axcell: Automatic extraction of results from machine learning papers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8580–8594.

S. Teufel, et al., Argumentative zoning: Information extraction from scientific text, Ph.D. thesis, Citeseer, 1999.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 2. Team L3S (Kabongo et al., 2024) | Leibniz University, Hannover, Germany
    a. Submitted a total of 12 runs for the few-shot evaluation phase 1 and 12 runs for the zero-shot evaluation phase 2.

References
S. Kabongo, J. D’Souza, S. Auer, Exploring the latest llms for leaderboard extraction, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Participant Approaches

- Participant 2. Team L3S (Kabongo et al., 2024) | Leibniz University, Hannover, Germany
  a. Submitted a total of 12 runs for the few-shot evaluation phase 1 and 12 runs for the zero-shot evaluation phase 2.
  b. **Approach:** Finetuned LLMs inspired after the FLAN-T5 strategy (Chung et al., 2024) which encompassed fine-tuning a pre-trained LLM with a standard set of instructions to better equip them to handle various tasks.
    - 4 models - Finetuned Mistral-7B and LLaMA 2 to make them better suited to handle the (T, D, M, S) extraction task. Furthermore, they also tested the most recent proprietary GPT models viz. GPT-4 and GPT-4o out-of-the-box.
    - 3 contexts (Kabongo et al., 2024) - As the information extraction context they tried 3 different methods: DocTAET ((T)-title, (A)- abstract, (E)-experimental setup, and (T)-tabular information parts of the full-text), DocREC (text selected from the sections named (R)-results, (E)-experiments, and (C)-conclusions), and DocFULL (full paper text).
    - Thus for each evaluation phase they submitted a total of 4 models x 3 contexts = 12 runs.

References

S. Kabongo, J. D’Souza, S. Auer, Exploring the latest llms for leaderboard extraction, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53.

S. Kabongo, J. D'Souza and S. Auer, Effective Context Selection in LLM-based Leaderboard Generation: An Empirical Study. In 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25–27, 2024. Springer LNCS 14762 and 14763.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

-

Reference

D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

Shows the binary classification performance.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

Shows the extraction performance w.r.t. ROUGE.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- Team AMATU's few-shot performance is w.r.t. AxCell

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5

Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- Team AMATU's few-shot performance is w.r.t. AxCell.
- And their zero-shot performance is w.r.t. GPT-3.5.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications
Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- Team L3S's results are w.r.t. the finetuned Mistral model.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- We see the finetuned model outperforms the rule-based or GPT model out-of-the-box

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Binary Classification and Extraction Performance w.r.t. the Rouge Summarization Metrics

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					Zero-shot				
	Rouge				Gen.	Rouge				Gen.
	1	2	L	Lsum	Acc.	1	2	L	Lsum	Acc.
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

- Nevertheless, Team AMATU presents novel insights into the community to leveraging LLM's effectively for the (T, D, M, S) extraction objective using clever prompt engineering strategies that shows comparable performance to computationally intensive finetuning approach

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Per (T,D,M,S) Element Extraction Performance w.r.t. the F-score.

Reference

D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Per (T,D,M,S) Element Extraction Performance w.r.t. the F-score.

Table 6
Evaluation results w.r.t. the individual (Task, Dataset, Metric, Score) elements and Overall in terms of **F1 score**. *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

		Few-shot					Zero-shot				
Model	**Mode**	**T**	**D**	**M**	**S**	**Overall**	**T**	**D**	**M**	**S**	**Overall**
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

- Zero-shot evaluations are lower than few-shot evaluations.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Per (T,D,M,S) Element Extraction Performance w.r.t. the F-score.

Table 6
Evaluation results w.r.t. the individual (Task, Dataset, Metric, Score) elements and Overall in terms of **F1 score**. *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

		Few-shot					Zero-shot				
Model	**Mode**	**T**	**D**	**M**	**S**	**Overall**	**T**	**D**	**M**	**S**	**Overall**
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

- Rule-based AxCell outperforms the finetuned LLM w.r.t. exact-match evaluations. However AxCell operates on a supplied taxonomy of known (T,D,M) whereas the finetuned models is generating (T,D,M) annotations without a supplied taxonomy.

Reference
D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Results

- Per (T,D,M,S) Element Extraction Performance w.r.t. the F-score.

Table 6
Evaluation results w.r.t. the individual (Task, Dataset, Metric, Score) elements and Overall in terms of **F1 score**. *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

		Few-shot					Zero-shot				
Model	**Mode**	**T**	**D**	**M**	**S**	**Overall**	**T**	**D**	**M**	**S**	**Overall**
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

- And among the T, D, M, and S extraction targets, the score element is the most challenging to extract.

Reference

D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

SOTA?: Tracking the State-of-the-Art in Scholarly Publications

Conclusions

- Our main findings are as follows:
  a. First, effective prompting paradigms should be a go-to strategy to test LLMs out-of-the-box for the SOTA? shared task objective.
  b. Second, finetuning small-scale models makes them better able to handle the SOTA? objective than larger-scale LLMs known for their generative AI abilities when simply applied to the IE task.
  c. Third, the paper context over which the IE task is expected to be performed must have an ideal balance of length versus selectivity of specific sections in the paper that indeed are highly likely to contain mentions of the (T, D, M, S). On the extreme end of the spectrum, using the full paper text without effective context selection hinders and seems to distract the LLM downstream IE task performance.

Reference

D'Souza, J., Kabongo, S., Babaei Giglou, H., & Zhang, Y. (2024). Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the state-of-the-art in scholarly publications. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum* (CEUR Workshop Proceedings). CEUR-WS. Online.

Funding Acknowledgement

The “SOTA?” track as Task 4 within the SimpleText 2024 evaluation lab at CLEF 2024 has been jointly funded by:

- The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259), and
- The German BMBF project SCINEXT (01IS22070).

CLEF 2024 GRENOBLE

TURN IT SIMPLE

TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Thank you for your attention!

Related Links:

- “SOTA?” Task 4 website: https://sites.google.com/view/simpletext-sota/home
- “SOTA?” Task 4 Codalab Competition Site: https://codalab.lisn.upsaclay.fr/competitions/16616
- “SOTA?” Task 4 Dataset: https://github.com/jd-coderepos/sota/

Discussions

Leibniz Association