REACTION MINER: An Integrated System for Chemical Reaction Extraction from Textual Data

Ming Zhong Siru Ouyang Yizhu Jiao Priyanka Kargupta
Leo Luo Yanzhen Shen Bobby Zhou Xianrui Zhong Xuan Liu
Hongxiang Li Jinfeng Xiao Minhao Jiang Vivian Hu Xuan Wang
Heng Ji Martin Burke Huimin Zhao Jiawei Han
University of Illinois Urbana-Champaign
{mingz5, siruo2, yizhuj2, pk36, hanj}@illinois.edu

Abstract

Chemical reactions, as a core entity in the realm of chemistry, hold crucial implications in diverse areas ranging from hands-on laboratory research to advanced computational drug design. Despite a burgeoning interest in employing NLP techniques to extract these reactions, aligning this task with the real-world requirements of chemistry practitioners remains an ongoing challenge. In this paper, we present REACTION MINER, a system specifically designed to interact with raw scientific literature, delivering precise and more informative chemical reactions. Going beyond mere extraction, REACTION MINER integrates a holistic workflow: it accepts PDF files as input, bypassing the need for pre-processing and bolstering user accessibility. Subsequently, a text segmentation module ensures that the refined text encapsulates complete chemical reactions, augmenting the accuracy of extraction. Moreover, REACTION MINER broadens the scope of existing pre-defined reaction roles, including vital attributes previously neglected, thereby offering a more comprehensive depiction of chemical reactions. Evaluations conducted by chemistry domain users highlight the efficacy of each module in our system, demonstrating REACTION MINER as a powerful tool in this field[1].

1 Introduction

Chemical reactions lie at the heart of chemistry, representing the transformative processes that give birth to new substances. The structured format of these reactions paves the way for diverse applications, including synthesis planning (Segler et al., 2018; Genheden et al., 2020), reaction prediction (Schwaller et al., 2018; Coley et al., 2019), and reaction condition recommendation (Gao et al., 2018; Maser et al., 2021). In recent years, the combination of chemistry and NLP has emerged as a dynamic research area (Chithrananda et al., 2020; Edwards et al., 2022; Bran et al., 2023), fueled by the prospect of automating the extraction of chemical reactions from vast corpora of scientific papers (Guo et al., 2022; Zhong et al., 2023). By leveraging NLP techniques, researchers can derive crucial insights more rapidly than traditional manual methods (Goodman, 2009), thereby catalyzing progress in myriad chemistry-related domains.

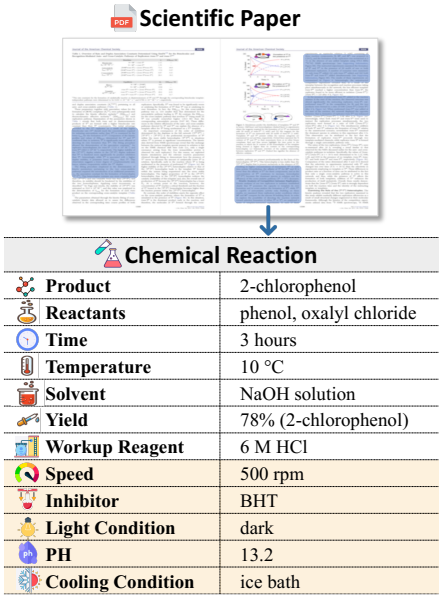

Figure 1: Example from REACTION MINER. Highlighted reaction roles denote the attributes that the previous systems are incapable of extracting.

Figure 1 illustrates the goal of this task, which is to mine and extract structured chemical reactions from the extensive chemical literature. Representatively, systems such as OPSIN, CHEMRXNBERT, and REACTIE have emerged to automate the process of chemical reaction extraction, each employing distinct NLP approaches. OPSIN (Lowe, 2012) serves as an early exemplar, utilizing a heuristic-based method that underscores the potential bene-

[1]Code, data, and models can be found at: https://github.com/maszhongming/ReactionMiner. The link to the video that introduces our system is at: https://youtu.be/q7P6NWDKcxw

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 389–402
December 6-10, 2023

fits of applying NLP to the realm of chemical data extraction. Subsequently, CHEMRXNBERT (Guo et al., 2022) harnesses the power of pre-training on chemistry literature to foster a deeper understanding of chemical context. REACTIE (Zhong et al., 2023) further refines the Information Extraction (IE) process by reformulating it as a Question Answering (QA) task, facilitating the creation of synthetic data and reducing annotation needs. Despite these significant strides, certain issues persist when these systems are deployed in the hands of real-world chemistry practitioners:

(1) **Input Format Misalignment**: Existing systems are designed to accept plain text as input. However, chemistry practitioners typically engage with literature in PDF format rather than processed text. This disconnect between the format of readily accessible resources and the input requirements introduces a considerable hurdle for practical use.

(2) **Limited Input Granularity**: The typical input for existing systems is confined to a sentence or a fixed-size context window for the extraction. This often leads to a trade-off between over- and under-inclusion of data, with systems either capturing incomplete information about the chemical reaction or introducing irrelevant content.

(3) **Restriction on Extracted Roles**: The current systems focus on extracting pre-defined reaction roles, such as the *reactant*, *product*, *catalyst*, *time*, *temperature*, *yield*, etc. Yet, there are additional attributes that are of considerable interest to practitioners, such as the *experimental procedure* and more nuanced *reaction conditions*, which are commonly overlooked by these systems.

(4) **Output Format Inconsistency**: Lastly, there exists a discrepancy between the output format provided by current systems and what is required by real-world users. Frequently, these systems output incomplete chemical names or incorrect symbols and units, which further complicates the interpretation and application of the extracted information.

To address the outlined challenges, we present REACTION MINER, an integrated system designed to bridge these gaps and cater more closely to the needs of real-world chemistry practitioners. In contrast to existing systems, REACTION MINER incorporates a series of new features:

(1) **PDF-to-Text**: Recognizing that the inherent diversity of templates in chemistry journals presents a formidable challenge for existing tools, we develop a PDF-to-text module specifically tailored for the biochemistry field. It features built-in dynamic similarity calculation functionality based on Sentence-BERT (Reimers and Gurevych, 2019) to alleviate the frequent coherence issues that arise during conversion.

(2) **Text Segmentation**: To ensure the input context includes all necessary details without irrelevant information about chemical reactions, we initially perform text segmentation on the processed text. This involves identifying the central sentence associated with the chemical reaction and subsequently expanding the boundaries of the input text using unsupervised topic segmentation (Choi, 2000).

(3) **Role Enrichment**: To bypass the restrictions imposed by pre-defined label space, we integrate an automatic event mining approach (Jiao et al., 2022) to enrich extracted reaction roles. Furthermore, to enhance our system's ability to accurately extract newly discovered attributes, we generate synthetic data corresponding to each role based on GPT-4 (OpenAI, 2023) for the training process.

(4) **Unified Reaction Extraction**: Striving to align our system's output more closely with the requirements of users, we unify the format of existing data and adjust the annotation guideline based on feedback from chemistry practitioners. Concretely, we re-collect (Zhong et al., 2023), re-organize[2], and re-annotate (Guo et al., 2022) present data, leading to a unified system that caters to the needs of the chemistry community more effectively.

Regarding evaluation, we invite chemistry Ph.D. students to undertake tests and contrast REACTION MINER with current systems. Human evaluation indicates that our system is better aligned with the needs of the chemistry community. Remarkably, even though the architecture of REACTION MINER is built upon LLaMA-7B (Touvron et al., 2023) and LoRA (Hu et al., 2022), it consistently matches or surpasses the performance of large language models across all subtasks. Thus, REACTION MINER represents a step forward in chemical reaction extraction, providing an accessible, high-performing open-source tool to expedite advancements at the intersection of NLP and chemistry.

2 Method

In this section, we start with the task formulation to provide an overarching perspective of REACTION MINER, subsequently delving into each module.

[2]Data from https://docs.open-reaction-database.org/en/latest.

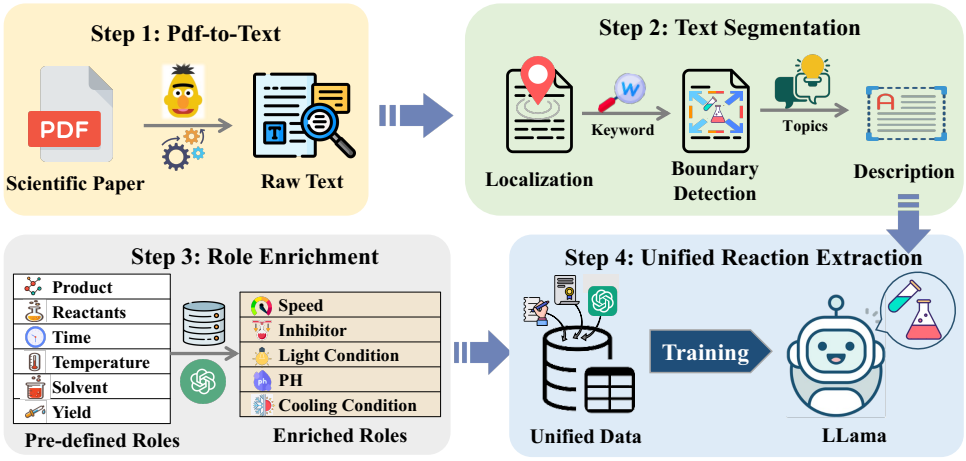

Figure 2: Overall framework for REACTION MINER.

2.1 Task Formulation

Given any PDF file that contains chemical context, the PDF-to-text module initially converts it into processed text, represented as T. This is followed by the segmentation model identifying and segmenting T into k relevant passages $\mathcal{P} = \{P_1, \cdots, P_k\}$ associated with the chemical reaction to serve as inputs for reaction extraction. For each passage P, the objective is to extract all the structured chemical reactions $\mathcal{C}$ present within P, where every reaction $C \in \mathcal{C}$ comprises n role-argument pairs $\{(r_1, a_1), \cdots, (r_n, a_n)\}$. Here, the term "role" r refers to a crucial attribute of a chemical reaction, such as the *product*, *reactant*, *catalyst*, *solvent*, *time*, *temperature*, *yield*, etc., while the corresponding "argument" a is the extracted span of the corresponding reaction role in the input P.

2.2 PDF-to-Text

The common format in which practitioners access literature is PDF rather than processed text, making it a more appropriate input for an extraction system. However, the inherent diversity in templates used by various chemical journals presents a significant challenge for the development of reliable PDF conversion tools. Current popular methods, such as Gorbid [3], which is used in S2ORC (Lo et al., 2020), and SymbolScrapper[4], either overlook short paragraphs or pose incoherence problems (such as intermixing header, footer, or caption information with the body text). These issues can ultimately impact the performance of subsequent extraction.

To address these challenges, we devise our own PDF-to-Text parser. It operates in three stages: 1) converting the given PDF files into XML format via SymbolScraper, 2) parsing the XML file into content paragraphs while excluding figures, tables, and captions from the body text through regular expressions, and 3) filtering out incoherent and irrelevant information (e.g., headers, footers, and references). This last step leverages the representative power of a pre-trained language model: we dynamically maintain a set of paragraphs representing the main content of the preceding paragraph and use the average embedding obtained from Sentence-BERT (Reimers and Gurevych, 2019) as the current anchor embedding. Subsequent paragraphs are filtered out if their cosine similarity to the anchor embedding falls below a certain threshold.

2.3 Text Segmentation

Text segmentation aims to segment out the reaction-related context from an entire paper. Segmenting the text into more manageable units enables the precise and efficient identification of reaction components. Its process is comprised of two main steps:

Keyword-based central sentence localization. The first step in the process involves leveraging the fundamental definition of chemical reactions, with the underlying hypothesis that all chemical reactions involve specific products (Muller, 1994). Human readers often intuitively identify products within textual information through linguistic cues, such as specific keywords. Consider the sentence, *"Removal of the TBS moiety of 17 was carried out with TBAF/AcOH in MeCN at 60 °C to give diol 20 in 86% yield."* From the presence of the word *"yield"*, one can deduce that the product is *"diol 20"*. Building on this insight, we curate a set of 35 keywords that are demonstrative of products in chem-

[3] https://github.com/kermitt2/grobid
[4] https://github.com/zanibbi/SymbolScraper

ical reactions. When scanning through the text, sentences containing these keywords are identified as central to understanding the reaction context.

Topic-aided boundary detection. Once the central sentence is identified, the subsequent step involves detecting the contextual boundary related to the specific chemical reaction. Research has shown that the integration of semantic information through topic models substantially enhances the effectiveness of segmentation algorithms (Riedl and Biemann, 2012; Alemi and Ginsparg, 2015). Motivated by these findings, we employ semantic topical information within the texts to more accurately discern different context blocks associated with various chemical reactions. Specifically, C99 (Choi, 2000), a widely-recognized method for topical detection is used. It annotates sentences with matching tags if they pertain to the same topical group. Any topical group containing the identified central sentence is considered as a segmented context, relevant to particular chemical reactions.

2.4 Role Enrichment

Typically, prevalent systems can extract 9 reaction roles, namely *product*, *reactant*, *catalyst*, *solvent*, *workup reagent*, *reaction type*, *time*, *temperature*, and *yield*. However, this coverage is insufficient for capturing all vital properties of a chemical reaction. To address this, we apply an event mining approach (Jiao et al., 2022) to the chemistry literature. It is grounded in the identification of all entities within a text, and then allowing T5 (Raffel et al., 2020) to generate the corresponding entity type to discover frequent new reaction roles. Upon manual review and filtering by chemistry practitioners, we integrate an additional 10 new reaction roles, with the complete list available in the Appendix C.

Simultaneously, an obstacle arises with the enrichment of reaction roles due to the current scarcity of suitable training data. To tackle this issue, we annotate descriptions of the newly added reaction roles and provide three demonstrations. This enables in-context learning, allowing GPT-4 to generate chemical text, alongside the corresponding extracted chemical reactions. We then institute a filtering process where we: 1) eliminate samples where the generated argument does not exist in the original text, 2) remove the generated roles does not appear in the label space, and 3) in tandem with REACTIE (Zhong et al., 2023), remove samples where the generated products exhibit low probabilities of extraction in REACTIE. As a result, we manage to enrich the spectrum of new roles that need extraction for the chemical reaction extraction task, coupled with the associated training data.

2.5 Reaction Extraction

Despite the existence of datasets for the chemical reaction extraction task, they vary in terms of reaction roles, output formats, and annotation guidelines. This underscores the need for a standardized and unified data format, which stands as a critical prerequisite for the development of a universally applicable system. Accordingly, we establish this requisite uniformity by re-collecting, re-organizing, and re-annotating the existing data as follows.

Re-collecting Negative Samples. While a keyword-based approach in the segmentation module currently serves to locate chemical reactions, this method is high in recall but low in precision. That is, passages containing the designated keywords do not necessarily describe chemical reactions. Thus, we incorporate negative samples — instances where the input text does not contain chemical reactions — into the reaction extraction training. For these samples, models should output "No complete chemical reaction". We achieve this by running our segmentation model on the chemistry literature and re-collecting the filtered segments as negative samples.

Re-organizing Open Reaction Database. The Open Reaction Database[5], a publicly available repository of chemical reactions, primarily contains data from patent literature, with ground truths extracted via the rule-based system OPSIN (Lowe, 2012, 2018). We re-organize the data format within this database, filtering out samples with semantically repetitive content in the input text. Additionally, our sampling procedure prioritizes scientific papers and examples containing multiple chemical reactions, forming part of our final training data.

Re-annotating Reaction Corpus. Although the Reaction Corpus (Guo et al., 2022) is a manually annotated dataset, its annotation guideline prompts the output chemical to be represented as a unique token of compound rather than the full name, reducing user readability. Moreover, its output format occasionally contains incorrect symbols and units due to tokenization errors. Thus, we re-annotate the

[5] https://docs.open-reaction-database.org

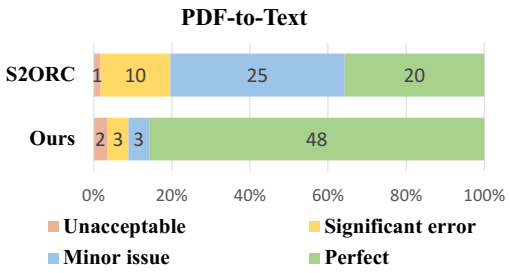

Figure 3: Human evaluation for PDF-to-Text.

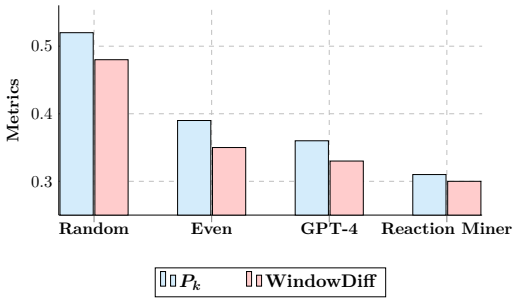

Figure 4: Evaluation results for text segmentation. Lower values indicate better performance.

training set to eliminate such minor inconsistencies, thereby achieving a unified data format.

Utilizing the above data, we train the LLaMA-7b (Touvron et al., 2023) in a parameter-efficient manner using the LoRA method (Hu et al., 2022), enabling it to function as a chemical reaction extractor within our REACTION MINER framework. More details can be found in the Appendix D.

3 Experiments

In this section, we evaluate the performance of REACTION MINER by testing its three core modules.

3.1 PDF-to-Text

Experimental Setup. To evaluate the quality and generalization of our PDF-to-Text parser, we randomly sample 56 papers from the top 10 most influential chemical journals across various scholarly publishers (i.e., The Royal Society of Chemistry and American Chemistry Society). For each sample, we manually compare the resulting text with the original PDF using a four-level rating system: perfect, minor issue, significant error, and unacceptable. Here, "minor issue" indicates a few incoherent lines, whereas "significant error" refers to an omission or mixture of several paragraphs, significantly impacting readability.

Results. The results of the human evaluation are shown in Figure 3. S2ORC (Lo et al., 2020) is a vast corpus of 81.1M English-language academic papers, thus its PDF-to-Text tool is widely employed. However, given the wide variety of journal templates in the chemical literature, it achieves a "perfect" rating in 20 instances, implying it only completely preserves 35.7% of the original text from the PDFs. Moreover, it frequently overlooks paragraphs or includes unrelated content. Conversely, the PDF-to-Text component in REACTION MINER flawlessly processes the text in 85.7% of the instances, underscoring the effectiveness of our proposed dynamic similarity computation component in resolving the incoherence issue.

3.2 Text Segmentation

Experimental Setup. We randomly collect 50 samples from chemical literature and employ 3 graduate students with chemical backgrounds to annotate the reaction-related context from the text. The average length of samples is 328.26, resembling a short article, and the average number of segments is 2.34 per sample. Three segmentation baselines are used for comparison: 1) `Random`: segment boundaries are randomly assigned; 2) `Even`: segment boundaries are evenly placed every k sentences; 3) `GPT-4`: employ GPT-4 to identify sentences related to chemical reactions. Two common measures for text segmentation, P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002) are leveraged as evaluation metrics, with lower values indicating better performance. During experiments, we set the size of the sliding window in WindowDiff $k = 2$, equaling k in `Even` baseline.

Results. Figure 4 presents a comprehensive summary of the text segmentation results obtained using the constructed test set. The analysis reveals that our proposed method substantially surpasses all the existing baseline methods with respect to both P_k and WindowDiff metrics, underlining its ability to accurately identify segment boundaries. A particular highlight of our findings is the superiority of our approach, REACTION MINER, over the strongest proprietary model, `GPT-4`, by improvements of 16.2% (0.37 $\Rightarrow$ 0.31) and 9.1% (0.33 $\Rightarrow$ 0.30) for P_k and WindowDiff, respectively. This not only establishes the efficacy of our method but also provides critical insights into the existing limitations of `GPT-4`, one of the most sophisticated language models to date, in the realm of text segmentation. Interestingly, `GPT-4`'s performance is found to be comparable to the `Even` method, point-

Input Text: It is worth mentioning that when the reaction of 2-cyano-N,N-dimethylethanethioamide (1s) with 3-phenylpropiolaldehyde (2a) was carried out at room temperature under nitrogen for 5 h, an aldol condensation product 5a was isolated in 73% yield, which can be further converted to 2-aminothienyl ether 3sa in 85% yield under the optimized reaction conditions.

ReactIE (Significant Error)	**GPT-4 (Minor Issue)**	**Reaction Miner (Perfect)**
Reaction 1:	Reaction 1	Reaction 1
Product: 5a	Product: aldol condensation product 5a	Product: 5a
Reactants: (1s), (2a), …	Reactant: 2-cyano-… (1s)	Reactant: 2-cyano-… (1s), 3-… (2a)
Reaction type: aldol condensation	Reactant: 3-… (2a)	Atmosphere: nitrogen
Catalyst: nitrogen	Atmosphere: nitrogen	Reaction type: aldol condensation
Temperature: room	Temperature: room temperature	Temperature: room temperature
Time: 5 h	Time: 5 h	Time: 5 h
Yield: 73%	Yield: 73% (aldol condensation product 5a)	Yield: 73% (5a)
Reaction 2:	Reaction 2	Reaction 2
Product: 3sa	Product: 2-aminothienyl ether 3sa	Product: 2-aminothienyl ether 3sa
Reactants: (1s), (2a), …	Reaction type: optimized reaction	Reactant: 5a
Yield: 85%	Yield: 85% (2-aminothienyl ether 3sa)	Yield: 85% (2-aminothienyl ether 3sa)

Figure 5: Case study for the different reaction extraction systems. The blue text indicates the results of the human evaluation, and the red text represents the parts of the model output that are problematic. In addition, the output of GPT-4 misses a few reaction roles, such as the "Reactant" in Reaction 2.

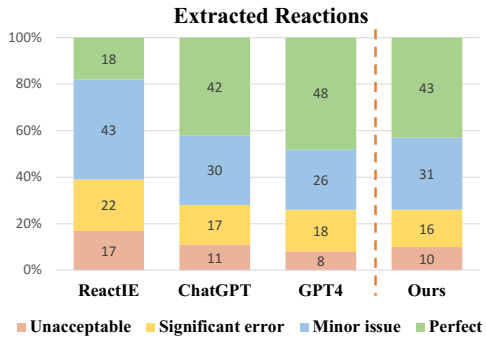

Figure 6: Human evaluation for reaction extraction.

ing to specific areas where even such a powerful model may exhibit weaknesses. Details of model outputs and implementation are in Appendix B.

3.3 Reaction Extraction

Experimental Setup. To exclusively assess the performance of the reaction extraction module, we curate a test set by manually annotating 100 samples that encompass the complete segment. We evaluate and contrast four distinct systems: 1) `ReactIE` (Zhong et al., 2023), which stands as the state-of-the-art chemical reaction extraction system, built upon Flan-T5 (Chung et al., 2022); 2) `ChatGPT`, a proprietary model for conversational scenarios based on InstructGPT (Ouyang et al., 2022); 3) `GPT-4` (OpenAI, 2023): the most advanced proprietary model accessible at present, and 4) Reaction extraction module in our REACTION MINER that utilizes LLaMA-7b as the backbone. Evaluation details are provided in Appendix D.

Results. The evaluation results are detailed in Figure 6. Despite being the previous best extraction system, `ReactIE` only perfectly matches the user's needs in 18% of the cases, with frequent minor issues and significant errors. Such shortcomings can be attributed to frequent formatting inconsistencies and the omission of certain reaction roles inherent in its prior data format. Contrastingly, the performance of REACTION MINER align more closely with `ChatGPT` and `GPT-4`, yielding a satisfactory outcome ("perfect" and "minor issue") in approximately 75% of cases. A salient point to highlight is that although having a considerably smaller parameter set than its proprietary counterparts and being open-source, REACTION MINER offers a performance that is on par. This positions it as a remarkably efficient open-source tool in this field. Figure 5 provides a more granular view of the outputs from different systems. In the given exemplar, `ReactIE` inaccurately identifies reactants and catalysts, resulting in it being categorized under "significant error". `GPT-4`, on the other hand, encounters a few formatting challenges, and misses reactants in the second reaction. In contrast, REACTION MINER adeptly extracts all the pertinent reaction roles, facilitating a comprehensive understanding of the given chemical reaction.

4 Conclusion

In our exploration, we present REACTION MINER, an integrated system adept at extracting chemical reactions directly from raw scientific PDFs. Beyond mere extraction, it offers enhanced accuracy by broadening the scope of reaction roles and eliminating prior gaps. Feedback from chemistry experts marks it as a powerful tool for the field.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

References

Alexander A. Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. *CoRR*, abs/1503.05543.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 26–33. ACL.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413. Association for Computational Linguistics.

Hanyu Gao, Thomas J Struble, Connor W Coley, Yuran Wang, William H Green, and Klavs F Jensen. 2018. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. 2020. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics*, 12(1):70.

Jonathan M. Goodman. 2009. Computer software review: Reaxys. *J. Chem. Inf. Model.*, 49(12):2897–2898.

Jiang Guo, A. Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W. Coley, Klavs F. Jensen, and Regina Barzilay. 2022. Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model.*, 62(9):2035–2045.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. 2022. Open-vocabulary argument role prediction for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5404–5418. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: the semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4969–4983. Association for Computational Linguistics.

Daniel Lowe. 2018. Chemical reactions from us patents (1976-sep2016). *doi*, 10:m9.

Daniel M. Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge, UK.

Michael R Maser, Alexander Y Cui, Serim Ryou, Travis J DeLano, Yisong Yue, and Sarah E Reisman. 2021. Multilabel classification models for the prediction of cross-coupling reaction conditions. *Journal of Chemical Information and Modeling*, 61(1):156–166.

P. Muller. 1994. Glossary of terms used in physical organic chemistry (iupac recommendations 1994). *Pure and Applied Chemistry*, 66(5):1077–1184.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguistics*, 28(1):19–36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557, Montréal, Canada. Association for Computational Linguistics.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. 2018. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098.

Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nat.*, 555(7698):604–610.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ming Zhong, Siru Ouyang, Minhao Jiang, Vivian Hu, Yizhu Jiao, Xuan Wang, and Jiawei Han. 2023. Reactie: Enhancing chemical reaction extraction with weak supervision. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12120–12130. Association for Computational Linguistics.

A Details for PDF-to-Text

Implementation Details. To filter out incoherent information, we dynamically maintain a set of anchor paragraphs. To avoid anchoring in the reference section, of which the embedding is totally different from other sections of a paper, we deliberately select the longest paragraph in the first one-third of the paper as the first anchor paragraph. To obtain embedding for each paragraph, we adopt a pre-trained sentence-transformer all-mpnet-base-v2[6]. Then we iterate through each paragraph, computing an average cosine similarity score between the paragraph embedding and each anchor paragraph. If the cosine similarity score falls below a threshold 0.12, we drop the content, otherwise, we add the current paragraph into anchor paragraphs. If the number of anchor paragraphs is more than 5, we pop the front-must anchor paragraph.

Evaluation Details. Papers used in evaluation are sampled from the following journals: Journal of American Chemistry Society, Angewandte Chemie International Edition, Chemical Communication, Chemical Society Reviews, Organic Letters, ACS Catalysis, The Journal of Organic Chemistry, Chemical Science, Organic & Biomolecular Chemistry, and Accounts of Chemical Research.

B Details for Text Segmentation

In this section, we provide supplementary information on the text segmentation module.

Keywords curation. All the following words are used as keywords when locating central sentences: *{ 'yields', 'yielded', 'yield', 'yielding', 'afforded', 'afford', 'affording', 'affords', 'produce', 'produces', 'produced', 'producing', 'obtained', 'obtain', 'obtaining', 'obtains', 'transformed', 'transform', 'transforms', 'transforming', 'convert', 'conversion', 'converted', 'converts', 'converting', 'synthesize', 'synthesized', 'synthesis', 'desired', 'desiring'}* Note that different words could have different forms of the same meaning.

Case Study. Figure 7 demonstrates the segmentation results from different models for one specific example in the test set. We can see that boundary relations predicted by GPT-4 are relatively simple compared to REACTION MINER, revealing its shortcomings in identifying contexts that are related to chemical reactions. REACTION MINER, on the other hand, provides a much similar boundary detection compared with the ground truth annotations.

[6] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

C Details for Role Enrichment

Reaction Role Definitions. Here we provide a complete set of 19 reaction roles as well as their definitions, including 9 roles prevalent in existing systems and another 10 roles enriched.

The following are the 9 reaction roles that are used in most existing systems:

(1) Product: Chemical substance that is the final outcome (major product) of the reaction.

(2) Reactant: Chemical substances that contribute heavy atoms to the product.

(3) Catalyst: Chemical substances that participate in the reaction but do not contribute heavy atoms (e.g., acid, base, metal complexes).

(4) Workup reagents: Chemical substances that are used after the reactions to terminate the reactions or obtain the products (e.g., quenching reagents, extraction solvent, neutralizing acids/bases).

(5) Solvent: Chemical substances that are used to dissolve/mix other chemicals, typically quantified by volume and used in superstoichiometric amounts (e.g., water, toluene, THF).

(6) Time: Duration of the reaction performed.

(7) Yield: Yield of the product.

(8) Reaction type: Descriptions about the type of chemical reaction.

(9) Temperature: Temperature at which the reaction occurs.

To capture sufficient information on chemical reactions and ensure coverage, we enrich the existing roles and obtain another 10 reaction roles mined from the chemistry literature. The definitions of enriched 10 reaction roles are listed below:

(1) Atmosphere: The type of gas present during the reaction can be crucial, especially for reactions sensitive to oxygen or moisture (e.g., reactions carried out under nitrogen or argon atmosphere).

(2) Inhibitor: Chemical substances introduced into the reaction environment to slow down, or completely halt, the reaction (e.g., a radical inhibitor like butylated hydroxytoluene (BHT) in polymerization reactions, a catalyst poison like sulfur in Haber process).

Ground Truth

When 0.5 equiv of Na2S2O8 was combined with 2 equiv of Selectfluor and 20 mol % AgNO3, the reaction time decreased from 2 h to 15 min for all substrates, forming alkyl fluorides in excellent yields. The significant acceleration of rate in the decarboxylative fluorination has led to a more efficient process, suggesting that this approach may be useful for 18F labeling. In the seminal work of Li, a Ag(II) fluoride is proposed as the active fluorine atom source in the reaction as opposed to Selectfluor. To support this supposition, Li and co-workers heated the combination of tert-butyl-2-ethyltetradecaneperox-oate and Selectfluor in a sealed tube to 120 \u00b0C for 2 h. When the reaction was run in acetone, a 22% yield of 3-fluoropentadecane was obtained, whereas when the reaction was run in 50:50 acetone/water, only a 4% of fluorinated product was obtained. On the basis of these findings, Li proposed that fluorine atom transfer from Selectfluor to alkyl radicals is unlikely to be involved in the Ag-catalyzed process. Selectfluor is reported to be unstable in water at high temperature, forming HF through reaction of the reagent and water. To examine this, we heated Selectfluor in acetone-d6/D2O to 120 \u00b0C in a sealed tube for 2 h. After cooling to room temperature, a sample was removed and examined by 1H NMR, showing that 80% of the reagent decomposed to the defluorinated chloromethyl derivative (see experiment were likely not conducive to testing whether radicals can abstract a fluorine atom from Selectfluor. In addition, if a Ag(II)-F intermediate was formed during the reaction, it can be present only in a catalytic amount (at most). During its formation, radicals are also generated in a catalytic amount, so the likelihood of a small amount of radical being fluorinated by a small amount of Ag(II)-F in the presence of excess Selectfluor is unlikely. Finally, there is a large body of evidence showing that Selectfluor and similar electrophilic fluorinating reagents react with radicals to form C-F bonds.

GPT-4 Segmentation

When 0.5 equiv of Na2S2O8 was combined with 2 equiv of Selectfluor and 20 mol % AgNO3, the reaction time decreased from 2 h to 15 min for all substrates, forming alkyl fluorides in excellent yields. The significant acceleration of rate in the decarboxylative fluorination has led to a more efficient process, suggesting that this approach may be useful for 18F labeling. In the seminal work of Li, a Ag(II) fluoride is proposed as the active fluorine atom source in the reaction as opposed to Selectfluor. To support this supposition, Li and co-workers heated the combination of tert-butyl-2-ethyltetradecaneperox-oate and Selectfluor in a sealed tube to 120 \u00b0C for 2 h. When the reaction was run in acetone, a 22% yield of 3-fluoropentadecane was obtained, whereas when the reaction was run in 50:50 acetone/water, only a 4% of fluorinated product was obtained. On the basis of these findings, Li proposed that fluorine atom transfer from Selectfluor to alkyl radicals is unlikely to be involved in the Ag-catalyzed process. Selectfluor is reported to be unstable in water at high temperature, forming HF through reaction of the reagent and water. To examine this, we heated Selectfluor in acetone-d6/D2O to 120 \u00b0C in a sealed tube for 2 h. After cooling to room temperature, a sample was removed and examined by 1H NMR, showing that 80% of the reagent decomposed to the defluorinated chloromethyl derivative (see experiment were likely not conducive to testing whether radicals can abstract a fluorine atom from Selectfluor. In addition, if a Ag(II)-F intermediate was formed during the reaction, it can be present only in a catalytic amount (at most). During its formation, radicals are also generated in a catalytic amount, so the likelihood of a small amount of radical being fluorinated by a small amount of Ag(II)-F in the presence of excess Selectfluor is unlikely. Finally, there is a large body of evidence showing that Selectfluor and similar electrophilic fluorinating reagents react with radicals to form C-F bonds.

Reaction Miner

When 0.5 equiv of Na2S2O8 was combined with 2 equiv of Selectfluor and 20 mol % AgNO3, the reaction time decreased from 2 h to 15 min for all substrates, forming alkyl fluorides in excellent yields. The significant acceleration of rate in the decarboxylative fluorination has led to a more efficient process, suggesting that this approach may be useful for 18F labeling. In the seminal work of Li, a Ag(II) fluoride is proposed as the active fluorine atom source in the reaction as opposed to Selectfluor. To support this supposition, Li and co-workers heated the combination of tert-butyl-2-ethyltetradecaneperox-oate and Selectfluor in a sealed tube to 120 \u00b0C for 2 h. When the reaction was run in acetone, a 22% yield of 3-fluoropentadecane was obtained, whereas when the reaction was run in 50:50 acetone/water, only a 4% of fluorinated product was obtained. On the basis of these findings, Li proposed that fluorine atom transfer from Selectfluor to alkyl radicals is unlikely to be involved in the Ag-catalyzed process. Selectfluor is reported to be unstable in water at high temperature, forming HF through reaction of the reagent and water. To examine this, we heated Selectfluor in acetone-d6/D2O to 120 \u00b0C in a sealed tube for 2 h. After cooling to room temperature, a sample was removed and examined by 1H NMR, showing that 80% of the reagent decomposed to the defluorinated chloromethyl derivative (see experiment were likely not conducive to testing whether radicals can abstract a fluorine atom from Selectfluor. In addition, if a Ag(II)-F intermediate was formed during the reaction, it can be present only in a catalytic amount (at most). During its formation, radicals are also generated in a catalytic amount, so the likelihood of a small amount of radical being fluorinated by a small amount of Ag(II)-F in the presence of excess Selectfluor is unlikely. Finally, there is a large body of evidence showing that Selectfluor and similar electrophilic fluorinating reagents react with radicals to form C-F bonds.

Figure 7: Case study for text segmentation conducted by the two most superior models. Note that the gray highlights are the boundary sentences.

(3) Pressure: The pressure at which the reaction is carried out, which may be above or below atmospheric pressure, depending on the requirements of the reaction.

(4) PH: If the reaction is carried out in an aqueous solution, the pH of the solution could be an important factor.

(5) Speed: Some reactions require specific stirring or mixing speeds, which can significantly impact the outcome of the reaction.

(6) Vacuum condition: Some reactions or post-reaction procedures (like solvent evaporation) require specific vacuum conditions to proceed effectively.

(7) Light condition: Certain reactions (photochemical reactions) require specific light conditions - wavelengths, intensity, or duration - to proceed.

(8) Cooling/Heating Condition: The specific conditions under which a reaction mixture is heated or cooled, including the temperature range, the rate of temperature change, and the duration at each temperature.

(9) Spectroscopic data: Information collected about the product using various spectroscopic methods such as NMR, IR, MS, which can help confirm its structure and composition.

(10) Procedure: The specific steps followed in conducting the reaction, including the order of addition of reactants, the sequence of reactions in multi-step syntheses, etc.

Source	# Texts	# Reactions
Open Reaction Database	200,000	201,666
GPT-4	35,173	48,305
Chemistry Literature	25,000	0
Reaction Corpus	385	491
Total	260,558	250,462

Table 1: Data statistics for reaction extraction. A text can contain more than one chemical reaction. "Chemistry Literature" is used as negative samples, i.e. the input text does not contain a chemical reaction.

Training Data Generation. To pair the newly enriched reaction roles with corresponding training data, we leverage GPT-4 with in-context learning technique for data generation. The detailed prompts are shown in Figure 8.

D Details for Reaction Extraction

Implementation Details. By re-collecting negative samples, re-organizing open reaction database, and re-annotating the reaction corpus, we gather the data statistics as presented in Table 1.

For the model training, we adopt the parameter-efficient approach LoRA to train the LLaMA-7b model as it is more computationally efficient and yields similar performance to full finetuning. The training is divided into two phases: we first conduct a two-epoch training on the reorganized open reaction database data and 22,000 negative samples from chemistry literature, expecting that the model can learn the preliminary chemistry knowledge in the first phase. Then, we perform the second stage of finetuning for a total of 20 epochs on the GPT-4 generated data, re-annotated reaction corpus, and 3,000 negative samples, aiming to allow the model to further learn to extract enriched reaction roles. For both phases of training, the batch size is set to 128, as well as the learning rate is 3e-4 with a warm-up ratio of 0.03.

Evaluation Details. For models belonging to the GPT class, we precede the input text with a comprehensive task definition and three exemplar demonstrations. Two Ph.D. students majoring in chemistry are invited to participate in manually evaluating the outputs of each model. The evaluation standards are consistently categorized into four tiers: perfect, minor issue, significant error, and irrelevant or invalid. The evaluation criteria continue to be divided into four levels: "perfect", "minor issue", "significant error" and "unacceptable".

The human evaluation system is built on an open-source annotation tool, doccano[7]. The evaluators coordinate the standards for the rating system before starting annotation and then each of them rate all the instances independently. They are presented with the input text and the model outputs. Model responses are listed in random order, with all the model information anonymized. Figure 9 provides a screenshot of the annotation interface. The reported performance in this paper is based on the results from all evaluators.

[7]https://github.com/doccano/doccano

Input Text: It is worth mentioning that when the reaction of 2-cyano-N,N-dimethylethanethioamide (1s) with 3-phenylpropiolaldehyde (2a) was carried out at room temperature under nitrogen for 5 h, an aldol condensation product 5a was isolated in 73% yield, which can be further converted to 2-aminothienyl ether 3sa in 85% yield under the optimized reaction conditions.
ReactIE (Significant Error) Reaction 1: Product: 5a Reactants: (1s), (2a), 2-cyano-N,N-dimethylethanethioamide, 3-phenylpropiolaldehyde Reaction type: aldol condensation Catalyst: nitrogen Temperature: room Time: 5 h Yield: 73% Reaction 2: Product: 3sa Reactants: (1s), (2a), 2-cyano-N,N-dimethylethanethioamide, 3-phenylpropiolaldehyde Yield: 85%
GPT-4 (Minor Issue) Reaction 1 Product: aldol condensation product 5a Reactant: 2-cyano-N,N-dimethylethanethioamide (1s) Reactant: 3-phenylpropiolaldehyde (2a) Atmosphere: nitrogen Temperature: room temperature Time: 5 h Yield: 73% (aldol condensation product 5a) Reaction 2 Product: 2-aminothienyl ether 3sa Reaction type: optimized reaction conditions Yield: 85% (2-aminothienyl ether 3sa)
REACTION MINER (Perfect) Reaction 1 Product: 5a Reactant: 2-cyano-N,N-dimethylethanethioamide (1s), 3-phenylpropiolaldehyde (2a) Atmosphere: nitrogen Reaction type: aldol condensation Temperature: room temperature Time: 5 h Yield: 73% (aldol condensation product 5a) Reaction 2 Product: 2-aminothienyl ether 3sa Reactant: 5a Yield: 85% (2-aminothienyl ether 3sa)

Table 2: Full text of case study in Figure 5. The blue text indicates the results of the human evaluation, and the red text represents the parts of the model output that are problematic. In addition, the output of GPT-4 misses a few reaction roles, such as the "Reactant" in Reaction 2.

Instruction

Please help me with a chemistry-related task which is divided into two steps: First, generate a paragraph in a scientific paper, which introduces one or multiple specific chemical reactions. Second, extract the information of all chemical reactions one by one from the generated paragraph. Completing these two steps generates an instance with paragraphs and a corresponding action list. Now please help me generate 5 instances.

Roles Definition

Specifically, each reaction should include several roles and their corresponding arguments. The roles are predefined attributes involved in the reaction while the arguments are the specific spans extracted from the paragraph that are corresponding to their roles. Here, we list all the reaction roles as below:

(1) Product: [def]	(2) Reactant: [def]	(3) Catalyst: [def]	(4) Workup reagents: [def]	(5) Solvent: [def]
(6) Atmosphere: [def]	(7) Inhibitor: [def]	(8) Reaction type: [def]	(9) Temperature: [def]	(10) Time: [def]
(11) Pressure: [def]	(12) PH: [def]	(13) Speed: [def]	(14) Vacuum condition: [def]	(15) Light condition: [def]
(16) Cooling/Heating Condition: [def]		(17) Spectroscopic data: [def]	(18) Yield: [def]	(19) Procedure: [def]

Output Format

For each instance, the output of our tasks should be in this format:
Paragraph:
[the generated text describing chemical reactions]

Reaction List:
Reaction 1
[the role-argument pairs of the first reaction. Note that the yield should be the its value followed by the corresponding product.]
...
Reaction n
[the role-argument pairs of the n-th reaction (if any)]

Demonstration (3 samples)

To clearly explain these tasks, we provide the following examples:

Instance 1
Paragraph:
Reactions of Zr derivatives such as 8 with [Ph3C][B(C6F5)4] in C6D5Br (at -20 and 20 °C, in absence and presence of d8-THF) were also performed. Although complicated mixtures were similarly produced, the generation of significant amounts of Ph3CCH2Ph were nonetheless observed (CH2 singlet appears at 4.04 ppm in C6D5Br [and 4.03 ppm in C6D5Br containing 3 drops of d8-THF]), suggestive of benzyl abstraction and benzyl cation formation (rather than trityl attack at the C(σ-aryl) atom) like that reported for the Zr-[O,N,C(σ-naphthyl)] analogue.

Reaction List:
Reaction 1
Product: Ph3CCH2Ph
Reactant: Zr derivatives such as 8, [Ph3C][B(C6F5)4]
Solvent: C6D5Br containing 3 drops of d8-THF
Reaction type: benzyl abstraction, benzyl cation formation
Temperature: -20 and 20 °C
Yield: significant amounts (Ph3CCH2Ph)

Instance 2
Paragraph:
Treatment of CyPBn-Cy with NiCl2(DME) in THF afforded (CyPBn-Cy)NiCl2, which is obtained as the dichloromethane solvate upon workup (68%) on the basis of elemental analysis, NMR spectroscopic, and X-ray crystallographic data. This material in turn was converted into (CyPBn-Cy)Ni(o-tol)Cl upon treatment with (o-tol)MgCl in THF and subsequently isolated as an analytically pure solid (95%).

Reaction List:
Reaction 1
Product: (CyPBn-Cy)NiCl2
Reactant: CyPBn-Cy, NiCl2(DME)
Solvent: THF
Yield: 68% ((CyPBn-Cy)NiCl2)

Reaction 2
Product: (CyPBn-Cy)Ni(o-tol)Cl
Reactant: (CyPBn-Cy)NiCl2, (o-tol)MgCl
Solvent: THF
Yield: 95% ((CyPBn-Cy)Ni(o-tol)Cl)

Instance 3
Paragraph:
1.44 ml (12.5 mmols) of benzoyl chloride, 1.88 ml (12.5 mmols) of N-benzyldimethylamine and 0.0281 g (0.125 mmol) of palladium acetate are added to 25 ml of toluene in a pressure apparatus constructed of glass. The apparatus is flushed with ethylene in order to remove the air. Ethylene is then injected at 10 bar and the mixture is stirred for 4 hours at 120° C. 55% of styrene and 9% of trans-stilbene are formed.

Reaction List:
Reaction 1
Product: styrene, trans-stilbene
Reactant: benzoyl chloride, N-benzyldimethylamine
Catalyst: palladium acetate
Solvent: toluene
Temperature: 120 °C
Time: 4 hour
Pressure: 10 bar
Yield: 55% (styrene), 9% (trans-stilbene)

Trigger

Following the above examples, please help me with this task, i.e. generate 5 instances containing paragraph and reaction list. Remember to strictly follow the output format. Ensure that all arguments of reactions are the real spans extracted from the paragraph and should be faithful to the original text. At least 1 of the 5 generated instances should contain multiple chemical reactions.

Figure 8: Prompt used for role enrichment. Specifically in "Role Definitions", [def] is the placeholder of definitions for all the reaction roles. For a complete list of definitions, please refer to Appendix C.

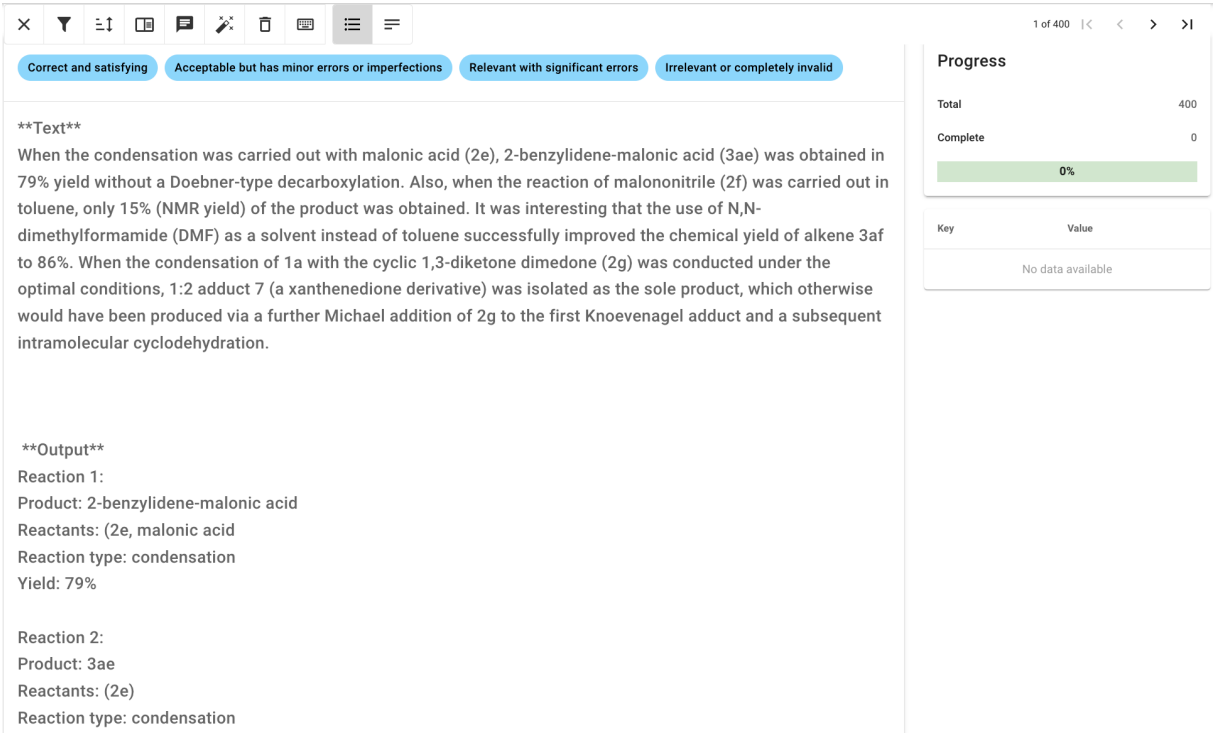

Figure 9: Annotation interface for human evaluation. The predictions from different models present in random order and the model information being anonymized. Our expert evaluators are required to read the input text, and then select the rating for the model's outputs from fours options for the extracted results.