Information Processing and Management 60 (2023) 103315

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

From "what" to "how": Extracting the Procedural Scientific Information Toward the Metric-optimization in AI

Yongqiang Ma [1], Jiawei Liu [1], Wei Lu, Qikai Cheng [*]

School of Information Management, Wuhan University, Wuhan, 430072, Hubei, China

ARTICLE INFO

Keywords:
Metric-driven mechanism
Information extraction
Scientific papers mining

ABSTRACT

In response to the exponential growth of the volume of scientific publications, researchers have proposed a multitude of information extraction methods for extracting entities and relations, such as task, dataset, metric, and method entities. However, the existing methods cannot directly provide readers with procedural scientific information that demonstrates the path to the problem's solution. From the perspective of applied science, we propose a novel schema for the applied AI community, namely a metric-driven mechanism schema **(Operation, Effect, Direction)**. Our schema depicts the procedural scientific information concerning "How to optimize the quantitative metrics for a specific task?" In this paper, we choose papers in the domain of NLP for our study, which is a representative branch of Artificial Intelligence (AI). Specifically, we first construct a dataset that covers the metric-driven mechanisms in single and multiple sentences. Then we propose a framework for extracting metric-driven mechanisms, which includes three sub-models: 1) a mechanism detection model, 2) a query-guided seq2seq mechanism extraction model, and 3) a task recognition model. Finally, a metric-driven mechanism knowledge graph, named MKG_{NLP}, is constructed. Our MKG_{NLP} has over 43K n-ary mechanism relations in the form of (Operation, Effect, Direction, Task). The human evaluation shows that the extracted metric-driven mechanisms in MKG_{NLP} achieve 81.4% accuracy. Our model also shows the potential for creating applications to assist applied AI scientists to solve specific problems.

1. Introduction

The number of academic papers in the domain of artificial intelligence (AI) has increased twelvefold in the last 20 years[2]. As a result, this "publications flood" has unfortunately buried a significant amount of valuable information. The extraction of structured information from unstructured scientific publications might reduce the amount of time that researchers must devote to information-seeking and thus improve the efficiency of research and development.

Lauriola et al., (2022) split NLP into two sub-branches: fundamental (or basic) research and applicative research. Similar to pure chemistry and applied chemistry, artificial intelligence can be divided into pure AI and applied AI. Moreover, pure AI scientists concentrate on exploring new models, algorithms, and theories at the cutting edge of fundamental AI problems. Applied AI scientists

[*] Corresponding author.
E-mail addresses: mayongqiang@whu.edu.cn (Y. Ma), laujames2017@whu.edu.cn (J. Liu), weilu@whu.edu.cn (W. Lu), chengqikai0806@163.com (Q. Cheng).

[1] Both authors contributed equally to this research.

[2] https://hai.stanford.edu/ai-index-2021.

https://doi.org/10.1016/j.ipm.2023.103315
Received 20 September 2022; Received in revised form 31 January 2023; Accepted 5 February 2023
Available online 12 March 2023

employ various artificial intelligence models, algorithms, and theories to achieve a specific practical goal or application.

The recent work on scientific information extraction in the AI domain has mainly focused on extracting domain-specific entities, such as tasks, datasets, metrics, and methods (D'Souza et al., 2020; Hou et al., 2021; Jain et al., 2020; P. Li et al., 2021; Luan et al., 2018; Zheng et al., 2021). These domain-specific entities represent the descriptive scientific information and answer factual questions beginning with the word "**What**," e.g., "*What is the issue addressed in this paper?*" and "*What dataset is adopted in this paper?*" Descriptive scientific information is predominantly geared toward pure AI scientists, assisting them to keep abreast with the state-of-the-art research. However, information needs vary from person to person. Applied AI scientists mainly need procedural scientific information, which should answer questions beginning with the word "**How**," such as "*How can we improve the accuracy of astronomical images classification?*" Descriptive scientific information is inadequate for assisting applied AI scientists to handle domain-specific problems, such as predicting 3D protein structures, imaging black holes, and automating drug discovery. Therefore, there exists a gap between the current scientific information extraction research and applied AI scientists' information needs. In this work, we focus on the information needs of applied AI scientists.

The mechanism reveals how to manipulate phenomena and assists individuals to comprehend the solution path (Glennan, 1996; Machamer et al., 2000). From the perspective of problem-solving, artificial intelligence can be viewed as the discovery and description of the mechanism that exists between a specific problem and its solution (McCarthy, 2007). Benchmarks, proposed to provide fair measurements of the research progress, have played a vital role in various AI-related problems. Achieving state-of-the-art performance on benchmarks is commonly regarded as a sign of advancement with regard to a specific problem. Therefore, optimizing the performance evaluation metrics using established benchmarks is a critical way to enhance the legitimacy of research (Koch et al., 2021). Following the benchmark-driven methodology (Schlangen, 2021), we define the metric-driven research pattern in AI as one that focuses on optimizing the performance of a specific task, as measured by the quantitative metrics.

Furthermore, the metric-driven mechanism described in our paper can be viewed as procedural scientific information on how to optimize the quantitative metrics of a specific task. Therefore, this paper primarily focuses on information concerning "how to optimize the specific measurable objective." The NLP domain was chosen for practice in this paper, since it is a representative, flourishing branch of AI. It is evident that extracting the procedural scientific information contained in scientific publications, particularly metric-optimization information, can improve the efficiency of searching, reading, and using.

As shown in Fig. 1, we construct a metric-driven mechanism representation schema to express the crucial procedural scientific information in AI. In our schema, the metric-driven mechanism is represented as triple in nature (***Operation***,***Effect***,***Direction***). Based on the proposed schema, we construct an annotated metric-driven mechanism extraction dataset from the abstracts of ACL papers[3]. We then propose a framework that adopts the pre-trained model (SciBERT (Beltagy et al., 2019) and BART(Lewis et al., 2020)) to extract metric-driven mechanism triples in single and multiple sentences. Finally, we construct a metric-driven mechanism knowledge graph in NLP, named MKG_{NLP}, to further improve the performance of knowledge retrieval.

In summary, our primary contributions are as follows:

- We propose a coarse-grained metric-driven mechanism representation schema. Furthermore, based on the proposed schema, an annotated dataset in the NLP domain is constructed, which contains 1,016 metric-driven mechanism triples, distributed across single or multiple sentences.
- We propose a mechanism extraction framework, which is composed of three sub-models: 1) a metric-driven mechanism detection model, 2) a query-guided seq2seq metric-driven mechanism extraction model, and 3) a task recognition model[4].
- A large number of publications from ACL are extracted to construct a metric-driven mechanism knowledge graph (KG) based on these trained models in our proposed framework. Human evaluations demonstrate that our metric-driven mechanism KG has a high degree of accuracy and utility.

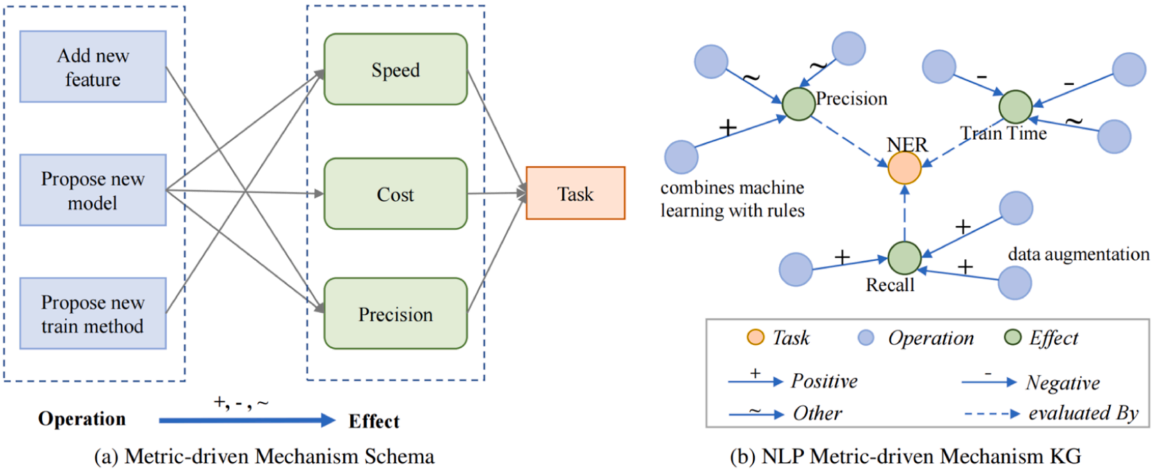
This article is organized as follows: Section II presents a brief literature review; Section III describes the research objectives; Section IV describes the schema and dataset of the metric-driven mechanism; Section V elaborates on the metric-driven mechanism's extraction framework; Section VI describes the experimental settings and provides insights into experimental results; and Section VII elaborates the construction of the metric-driven mechanism knowledge graph based on the proposed framework and further analyzes the application of the knowledge graph. The final section concludes this work and suggests directions for future work.

2. Literature Review

In this section, we will first discuss the definition and related work concerning the mechanism in the field of science, then further introduce and summarize the related works on scientific information extraction and scientific knowledge graphs. Finally, we will outline similarities and differences between our work and the existing works.

[3] https://aclanthology.org.

[4] The data and code are available at https://github.com/mayq97/metric_driven_mechanism.



(a) Metric-driven Mechanism Schema

(b) NLP Metric-driven Mechanism KG

Fig. 1. The Operation refers to a paper's innovative algorithm, model, and method. The Effect refers to the metrics evaluating the operation's efficiency and performance. The symbols +, − , ~ between operation and effect refer to the change Direction types of Effect entity. Here, we divide direction into three categories: positive, negative, and other.

2.1. The Mechanism in the Science Field

According to the Oxford English Dictionary, a mechanism is "a natural system or type of behavior that performs a particular function"[5]. In the philosophy of science, there is a great deal of discussion about the formal definition of the mechanism. For example, Machamer et al., (2000) defined mechanisms as organized entities and activities that produce regular changes from start or set-up to finish or termination conditions. Glennan, (1996), meanwhile, defined the mechanism as a complex system that produces behavior through the interaction between several parts, according to direct causal laws.

Mechanisms are involved in many research disciplines, and their definition varies across disciplines. In the field of biology, biochemists and molecular biologists pursue explanations of genes, proteins, and the molecules that influence biochemical reactions in the context of mechanistic explanations (Bechtel, 2010; Röhl, 2012; Yang et al., 1996). In the field of chemistry, researchers regard chemical reactions as a mechanism. In the field of social science, social mechanisms are considered complexes of agents' interactions, which underlie and account for macro social regularities (Steel, 2007). In the fields of AI, the mechanism can be viewed as a procedure that it applied to specific tasks or problems.

2.2. Scientific Information Extraction

Information Extraction (IE) refers to the extraction of structured information from unstructured or semi-structured texts. Information extraction from scientific literature enables researchers to obtain key insights from scientific documents. The current scientific information extraction ranges across the fields of computer science, biomedicine, and chemistry. Furthermore, AI is a flourishing branch of computer science.

In general, the problem of IE comprises Named Entity Recognition (NER) and Relation Extraction (RE) tasks. There are two types of approaches to IE: the pipeline-based and the end-to-end joint approaches. Regarding the former approach, the model first recognizes entities using a NER method, then extracts the relations between the recognized entities using an RE method (Chan & Roth, 2011; Lin et al., 2016; Zhong & Chen, 2021). The strength of the pipeline-based approach is its flexibility when integrating different data sources and methods, but its weakness is the error propagation problem between the individual NER and RE steps. Regarding the end-to-end joint approach, the model jointly extracts entities and relations using the shared layer or shared parameters between the NER and the RE tasks (Eberts & Ulges, 2020; D. Ji et al., 2020; Wadden et al., 2019; Zheng et al., 2021).

Moreover, X. Li et al., (2019) designed an alternative strategy in which they cast the entity-relation extraction as a multi-turn question-answering problem. Cui et al., (2021); Yan et al., (2021) proposed a unified sequence-to-sequence (Seq2Seq) framework based on BART (Lewis et al., 2020) to extract entities in the text for flat, nested, and discontinuous NER subtasks. X. Chen et al., (2022), meanwhile, proposed a lightweight generative framework with prompt-guided attention for low-resource NER.

In the case of scientific information extraction, the current scientific information extraction in the AI domain primarily focuses on extracting keyphrases (Y. Jiang et al., 2021; Kim et al., 2010), lexical functions of keyphrases (Lu et al., 2020), fine-grained scientific entities (e.g. *Task, Method, Dataset*, and *Metric*) (D'Souza et al., 2020; Hou et al., 2021; Jain et al., 2020; P. Li et al., 2021; Luan et al., 2018) and relations (Augenstein et al., 2017; Gábor et al., 2018; Mondal et al., 2021). In SemEval 2017 Task-10, Augenstein et al., (2017) proposed the *hyponym-of* and *synonym-of* relations. In SemEval 2018 Task-7, Gábor et al., (2018) proposed the *usage, result,*

[5] https://www.oxfordlearnersdictionaries.com/definition/academic/mechanism.

part-whole, and *compare* relations. Recently, Mondal et al., (2021) proposed the *evaluated-On* and *evaluated-By* relations.

2.3. Scientific Knowledge Graphs

Knowledge graphs (Fan & Wang, 2022; S. Ji et al., 2022) are a large linked semantic network of entities and relationships, that display huge potential for improving scientists' daily research efficiency. For example, Papers With Code (PWC) assists the researcher to search for and compare relevant works. Following the development of NLP, researchers have proposed many Scholarly Knowledge Graphs (SKGs) that are capable of serving specific user needs in the fields of AI. For example, Huo et al., (2022) leverage a novel dynamic Bibliographic Knowledge Graph (BKG) with a pre-trained node embedding to improve scientific topic hotness prediction performance. Dessì et al., (2020) proposed the Artificial Intelligence Knowledge Graph (AI-KG) to support researchers' daily work. AI-KG has five types of entities (tasks, methods, metrics, materials, and others) and 27 relations. Kabongo et al., (2021) constructed a Leaderboard for various AI subdomains to enable AI scientists to keep track of research progress. The AI-Leaderboard includes Task, Dataset, Metric (TDM) triples. Mondal et al., (2021) constructed an NLP TDM KG based on scientific papers using the end-to-end approach. In short, the current research emphasizes descriptive information (e.g., task, dataset, metric, and their interrelations) rather than procedural information in scientific publications.

2.4. Similarities and Differences with Existing Works

There exist several studies related to mechanism representation and extraction in different domains. Hope et al., (2017) proposed a weak structural representation that describes an idea using product descriptions regarding the purpose (*what they are trying to achieve*) and mechanism (*how they achieve that purpose*). V. Z. Chen et al., (2020) identified the hypotheses within scientific documents related to the fields of business and management, from which they then extracted the cause and effect entities in those hypothesis sentences. Hope et al., (2021) built a COVID-19 mechanism relations knowledge base, which includes activities, functions, and influences relations extracted from the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020). In summary, all of the previous constructed a very simple mechanism knowledge representation schema, which is an optimal solution considering the trade-offs among ease of extraction, scalability, and coverage. The main difference between our work and these previous studies is that we focus on the mechanism information concerning metric-optimization information in the fields of AI.

Scholars (D'Souza et al., 2020; Eberts & Ulges, 2020; Hou et al., 2021; Jain et al., 2020; P. Li et al., 2021; Zheng et al., 2021) proposed various methods and datasets in the fields of AI to solve the scientific information extraction problem. The main differences between these and our current work are as follows:

- The orientation of the extraction differs. The current work primarily focuses on descriptive scientific information, such as tasks, datasets, metrics, method entities, and entity relations. Our work shifts the attention from descriptive scientific information to procedural scientific information. We propose a metric-driven mechanism schema (Operation, Effect, Direction) to represent the metric-optimization information stated in the papers' abstracts.
- The information extraction mode differs. For scientific information extraction, researchers formalize it as a natural language understanding task in a sequence-labeling style. To extract the long entities in the metric-driven mechanism at the paragraph level, we transfer the metric-driven mechanism extraction from a sequence-labeling format into a multi-turn text generation format. To avoid ambiguous entities, we formalize the task entity extraction problem as a multi-label classification task.

3. Research Objectives

As described in the Introduction section, the current scientific information extraction methods primarily answer questions beginning with the word "**What**." The purpose of this study is to find procedural scientific information that answers questions beginning with the word "**How**." Specifically, our research objective is to extract the procedural scientific information from the abstract text to answer the typical research problems in AI (i.e., optimizing specific metrics for related tasks) and build a scientific knowledge graph for metric-optimization information in AI. Therefore, we propose to answer the following two research questions:

Question 1: how can we represent the metric-optimization information stated in the paper abstract?

Question 2: how can we extract the metric-optimization information?

For question 2, we divided the target problem into three subtasks: Subtask 1: Abstract selection. Identify abstracts containing a metric-driven mechanism; Subtask 2: Mechanism extraction. Extract the metric-driven mechanism from the recognized sentences; and Subtask 3: Task extraction. Map a given abstract to a predefined tasks' taxonomy.

4. The Metric-driven Mechanism Schema and Dataset

In this section, we will first introduce the metric-driven mechanism from the distribution, and schema to statement patterns in the abstract. Then, we will further propose the dataset for metric-driven mechanism extraction.

4.1. The Metric-driven Mechanism

4.1.1. Distribution of the Metric-driven Mechanism

A detailed mechanism description is required in many scientific fields in order to deliver a satisfactory explanation (Machamer et al., 2000). As shown in Table 1, the mechanism for specific measurable objectives exists in AI (e.g., natural language processing and computer vision), chemistry, biology, and other domains.

The abstracts of scientific research papers provide the readers with an informative summary of the entire paper. The scientific abstract is primarily composed of four parts: Introduction, Methods, Results, and Discussion. In the fields of AI, benchmarks (Martínez-Plumed et al., 2021) formalize a particular task through datasets and associated measurable metrics. To increase the legitimacy of research work, it is common to improve the metric value related to specific target tasks on an openly-established benchmark. As a result, researchers explicitly state that their models improve or reduce specific metrics in the paper's abstract in the AI field.

To analyze the distribution of metric-driven mechanisms, we manually annotated the long papers in ACL. To ensure the representativeness of the selected papers and comparability between different years, we choose papers in ACL 2020, and ACL 2021 as representative of the recent research in the NLP field, and papers in ACL 2010, and ACL 2011 as representative of earlier research in the NLP field. As shown in Table 2, about 40% of the papers' abstracts contain a mechanism for optimizing task-related metrics. Moreover, the percentage of metric-driven mechanisms has increased slightly in recent years.

Moreover, we randomly selected 100 papers from ACL for a fine-grained analysis of the distribution of metric-driven mechanisms in the NLP domain. We found that 76% of the papers contained a metric-driven mechanism, while a further 24% of them focus on research on constructing a new dataset, proposing a novel task, and conducting an empirical or theoretical analysis.

Further, we divide papers containing the metric-driven mechanism into two categories:

(1) Providing clear, specific metric-driven mechanism descriptions in the abstract, such as "using only half the number of parameters, our model achieves competitive accuracy with the best extractor, and is faster", accounts for 44% of the papers.
(2) Claiming the qualitative effect of the proposed method in the abstract, such as "we argue that discourse references have the potential of substantially improving textual entailment recognition", accounts for 32% of the papers. Although there are no specific metric-driven mechanism descriptions in the abstract, based on an analysis of partial full-text data, we found that the main sections (e.g., the results and discussion sections) of the papers contained many metric-driven mechanisms.

Due to the limited availability of full-text data, we mainly focus on the former category in this paper. We also have tested our model for the main sections of the papers. The results show that our model can extract the metric-driven mechanism from the main sections. In future, we will further explore the metric-driven mechanism extraction from the bodies of papers, once the papers' full-text data are available.

4.1.2. Schema for the Metric-driven Mechanism

The mechanism expresses the causal relationship between the phenomena or entities within a schema. These phenomena or entities that are expressed in the mechanism vary from the concrete (e.g., chemicals, cells, and plants) to the abstract (e.g., theories and concepts). There exist several studies related to mechanism extraction and representation (V. Z. Chen et al., 2020; Hope et al., 2021), all of which construct a straightforward mechanism representation schema. According to the characteristics of AI research, we divide these phenomena or entities into two types, *operation* and *effect*, based on the trade-offs among ease of extraction, scalability, and coverage.

Metrics play a central role in AI research, and are found metrics in 42% of the ACL papers' abstracts. The metric in the abstract is employed as a comparable indicator to state the effect of the proposed methods and models on specific tasks. In this work, we propose a metric-driven mechanism schema, in which the metric-driven mechanism in natural language is abstracted as a triple in nature $(\textbf{Operation}, \textbf{Effect}, \textbf{Direction})$: *Operation* represents an entity, such as a method or a model, proposed by the researcher; *Effect* refers to the metrics evaluating the operation's efficiency and performance, such as the F1-score, error rate, and time cost; and *Direction* expresses the relationship between the operation entity and the effect entity.

Effect is a measurable, comparable entity in a metric-driven mechanism schema. Therefore, we use the trisection method to divide the *Direction* in the metric-driven mechanism triple into *positive, negative*, and *other*.

- *Positive*: the method/model proposed in the research article improves the metric; for example, the pretraining model improves the F1 score of the text classification task.
- *Negative*: the method/model proposed in the research article reduces the metric; for example, using structural features to reduce the alignment error rate.
- *Other*: other than the above two relationships; for example, an external feature affects the metric, but we did not know the effect's direction.

4.1.3. Statement of the Metric-driven Mechanism

According to the range of metric-driven mechanism texts, the metric-driven mechanism in the abstract includes two expression types: 1) existing in a single sentence; and 2) existing in multiple sentences, as shown in Table 3. The forms of the metric-driven mechanism existing in a single sentence can be further divided into two types: explicit description and implicit description. Therefore, we focus on metric-driven mechanisms in both single and multiple sentences in this paper.

Table 1
The Mechanism in scientific research. The example mechanisms are drawn from scientific abstracts in the fields of natural language processing (NLP), computer vision (CV), chemistry (Chem), and biology (Bio).

No.	Example	Field
1	We apply **SVM ranking models** and achieve an exact sentence **accuracy** of 85.40% on the Redwoods corpus.	NLP
2	In this paper, we experimentally study the **combination of face and facial feature detectors** to improve **face detection performance**.	CV
3	The **rate of reduction** is decreased by **increasing amounts of stabilizing agents** and increased by increasing concentrations of precursor ions.	Chem
4	**Low light availability** and **high nutrient availability** increased the **nitrogen content of leaf** tissue by 53% and 40% respectively, resulting in a 37% and 31% decrease in the C/N ratio.	Chem
5	In conclusion, **high-energy diet** may improve **number of small follicles** and alter **energy metabolite** during early luteal phase in cycling ewes.	Bio
6	**Knocking down the expression of TaLSD1** through virus-induced gene silencing (VIGS) increased **wheat resistance** against Pst accompanied by an enhanced hypersensitive response (HR), an increase in PR1 gene expression and a reduction in Pst hyphal growth.	Bio

Table 2
Distribution of the metric-driven mechanism. About 40% of papers' abstracts contain one.

Year	Number of total abstracts	Number of abstracts containing metric-driven mechanism	Percentage
2010	159	58	36.5%
2011	162	73	45.1%
2020	778	315	40.5%
2021	571	255	44.7%

Table 3
Metric-driven mechanism statement patterns in the abstract.

Mechanism Position	Description	Example
Single sentence	Explicitly state the effect of the innovative model or method on the specific metric or aspect.	The structured neural parser achieves a 1.8% accuracy improvement.
	Implicitly state the effect of the innovative model or method on the specific metric or aspect by comparison	1. We compare LDA-SP to several state-of-the-art methods achieving an 85% increase in recall at 0.9 precision over mutual information (Erk, 2007). 2. We show that scaling to large topic spaces results in much more accurate models.
Multiple sentences	The effect entities and operation entities are separated from each other in different sentences.	For decoding, we describe a **coarse-to-fine approach** based on lattice dependency parsing of phrase lattices. We demonstrate **performance** improvements for Chinese-English and Urdu-English translation over a phrase-based baseline.

4.1.4. Differences with Existing Schemas

The current scientific information extraction schema, especially the SCIERC, primarily focuses on descriptive scientific information, such as tasks, metrics, method entities, and entity relations. As shown in Fig. 2, the schema proposed in SCIERC uses the *evaluate-for* to describe the relationship between the "*triphone and semi phone systems*" and *"error rate,"* which cannot represent information about metric optimization.

Our metric-driven mechanism schema shifts the attention from descriptive scientific information to procedural scientific information. We compared our mechanism recognition result with the schema in SCIERC. As shown in Fig. 2, using our schema, the information on metric optimization can be extracted; for example, the *"improved duration model"* has a negative effect on the *"error rate."* Based on the extracted metric-driven mechanism, questions regarding metric optimization can be answered, such as "How can we reduce the error rate of triphone and semi phone systems?"

4.2. Dataset

4.2.1. Dataset Construction

To the best of our knowledge, there is no available annotated dataset for metric-driven mechanism extraction. Therefore, we construct a metric-driven mechanism extraction dataset that primarily includes **mechanism identification** and **mechanism annotation**. In the dataset annotation stage, the annotated result is independent of the published paper's year.

In the **mechanism identification step**, the annotator must judge whether or not the abstract contains the metric-driven mechanism. Abstracts that do so are usually linked with the metric entities and cue verbs (e.g., improve, reduce) that reflect the change direction. Therefore, we use two heuristic rules to pre-label the ACL papers' abstracts to improve the annotation efficiency. Specifically, heuristic rules primarily consider two aspects: verbs (e.g., effect, influence, decrease, reduce, increase, and improve), and metric entities (e.g., accuracy, F1 score, BLEU, performance, and quality). The SpERT (Eberts & Ulges, 2020), trained on the SCIERC dataset (Luan et al., 2018), is employed for the metric entities recognition.

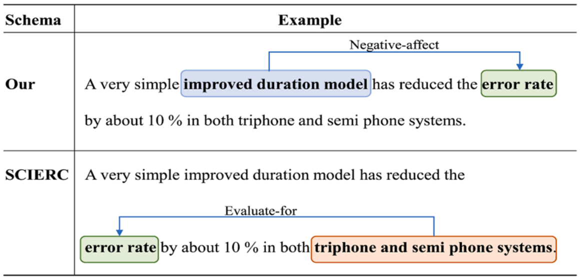

Schema
Example
Our
Negative-affect
A very simple improved duration model has reduced the error rate
by about 10 % in both triphone and semi phone systems.
SCIERC
A very simple improved duration model has reduced the
Evaluate-for
error rate by about 10 % in both triphone and semi phone systems.

Fig. 2. Comparison between our schema and SICERC. SICERC primarily focuses on descriptive information. Whereas our schema can represent information about metric-optimization.

In the **mechanism annotation step**, the annotator must label the effect and operation entities in the metric-driven mechanism's schema, then determine the relationship between the effect and operation entities based on context analysis and reasoning. Here, we employ BRAT[6] as the annotation tool for metric-driven mechanism tagging. The two annotators are graduate students with an NLP background. In cases of annotation disagreement on the entity boundaries, we choose the longer annotation.

For the metric-driven mechanism in a single sentence, some of the operation entities are pronoun phrases, such as *the proposed model, our model*, and *the model*. When pronoun phrases play the role of an operation entity in metric-driven mechanisms, the annotators are required to find and label the actual operation entity in the abstract context based on reasoning.

Moreover, we use the paper metadata in PWC as a training dataset for task extraction. PWC is an open-source repository of papers, datasets, and evaluations in the fields of machine learning and natural language processing, created by researchers at Facebook AI Research[7]. PWC contains a taxonomy of tasks and subtasks (Koch et al., 2021). There exist 2,328 task categories in the PWC dataset.

4.2.2. Annotated Dataset Analysis

Based on the annotated dataset, summaries of the statistics for the datasets are provided in Table 4 and Table 5. As shown in Table 4, the proportion of papers' abstracts that describe metric-driven mechanisms is 39.5%. We find that the operation entity is primarily distributed in the third to sixth sentences, and the effect entity is distributed mainly in the fourth to eighth sentences, as shown in Fig. 3a. In Table 5, the distribution of metric-driven mechanism relations is also highly imbalanced, with positive affect relations accounting for the majority. Approximately 55% of the metric-driven mechanisms exist in the same sentence. For those existing in multiple sentences, we find that the operation entity is usually stated before the effect entity, as shown in Fig. 3b, which is consistent with the expression pattern of the abstract; i.e., the IMRaD structure[8].

5. Methodology

Recently, the pre-trained models, e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2020) have promoted the performance of natural language understanding tasks ranging from text classification and named entity recognition to text generation. We propose a framework based on a pre-trained model to extract the metric-driven mechanisms. As shown in Fig. 4, our framework contains three components: 1) a metric-driven mechanism detection model, 2) a query-guided sequence-to-sequence (seq2seq) metric-driven mechanism extraction model, and 3) a task recognition model. In this work, the metric-driven mechanism detection model and extraction model are trained on the annotated dataset on ACL papers' abstracts (Section 4.2). For task recognition, we leverage the paper metadata in PWC as the model training dataset.

5.1. Mechanism Detection

We formalize the metric-driven mechanism detection as a binary text classification task. Given a scientific paper's abstract X_{abs}, the model must identify whether or not X_{abs} contains a complete metric-driven mechanism. SciBERT (Beltagy et al., 2019) is a pre-trained language model based on BERT for the scientific domain, which utilizes large-scale scientific publications as a pretraining task dataset and advances downstream scientific NLP tasks. Therefore, our metric-driven mechanism detection model uses SciBERT as the backbone for extracting the text's semantic information. Our BERT-based mechanism detection model contains two parts; i.e., a text encoder and a classification layer.

In the text encoder, we employ SciBERT to extract the text features of the input paper abstract. The input of the text encoder can be represented as follows:

$$\boldsymbol{X_{abs}} = [[CLS], token_1, token_2, \cdots, token_m, [SEP]] \tag{1}$$

[6] https://brat.nlplab.org/standoff.html.

[7] We downloaded the PWC dataset (licensed under CC BY-SA 4.0), and primarily focused on the Papers with abstracts archive.

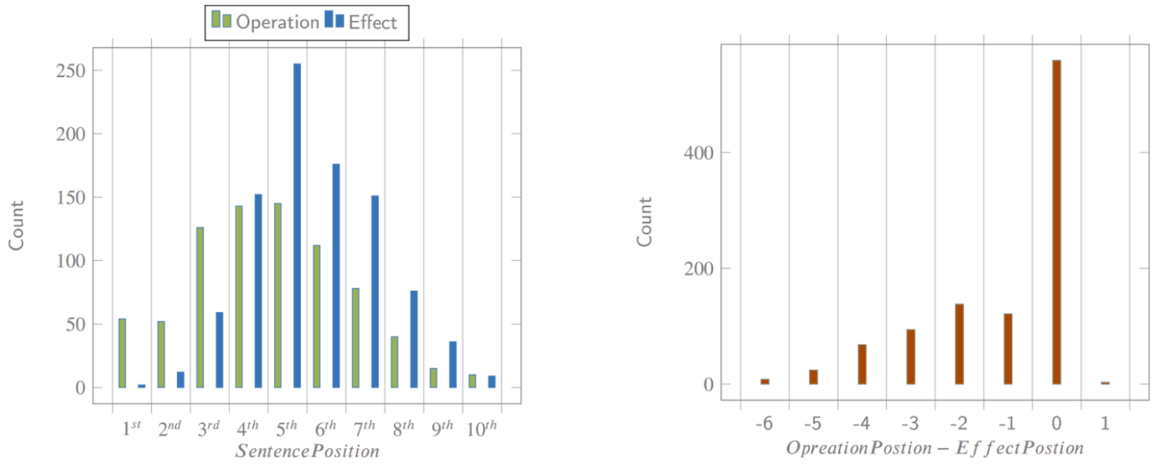
[8] https://en.wikipedia.org/wiki/IMRAD.

Table 4
Statistics for the dataset for subtask 1.

Statistics items	Number
Total # of abstracts	1,670
# of abstracts containing metric-driven mechanism	660
# of abstracts not containing a metric-driven mechanism	1,010

Table 5
Statistics for the dataset for subtask 2.

Type	Statics items	Number
Eentity	# of Entities	1,712
	# of Operation Entities	778
	Avg # of Operation Entity Tokens	4.36
	# of Effect Entities	934
	Avg # of Effect Entity Tokens	1.44
Relation	# of Total relations	1,016
	# of Positive	652
	# of Negative	274
	# of Other	90



(a) Absolute position distribution of metric-driven mechanism entities in abstracts.

(b) Relative position distribution of metric-driven mechanism entities in abstracts.

Fig. 3. Position distribution of metric-driven mechanism entities in abstracts. (a): The operation entity is in the first half of the abstract, and the effect entity in the second half. (b): The X-axis represents the distance between the operation entity and the effect entity. A negative value denotes that the operation entity is before the effect entity, and 0 denotes that the operation entity and effect entity exist in the same sentence.

where $token_i$ denotes the i^{th} token of the X_{abs} as tokenized by the corresponding tokenizer. m is the token number of X_{abs}, and $[CLS]$ and $[SEP]$ correspond to the special symbol at the abstract's beginning and end, respectively. We can obtain the text vector representation h via SciBERT through multiple Transformer (Vaswani et al., 2017) layers.

$$h = SciBERT(X) \tag{2}$$

In the classification layer, we use h_{CLS}, the first component of h, corresponding to the $[CLS]$ token, as the input to the classification layer. We use a fully connected layer in the classification layer to map the h_{CLS} to a 2-dimensional label logits vector, then a softmax function to the label logits to obtain the probability distribution regarding whether or not the input abstract contains a metric-driven mechanism.

$$p = Softmax(W \cdot h_{CLS} + b) \tag{3}$$

where p is a 2-dimensional vector that denotes the probability that the abstract contains mechanisms, and $\mathbf{W}$ and $\mathbf{b}$ denote the weight and bias in the fully connected layer, respectively.

Limited by the training data, we employ the data augmentation to enhance the model's performance. We add sentence granularity text to the original paragraph granularity training dataset. Specifically, if the sentence contained an effect entity, it was labeled 1. To

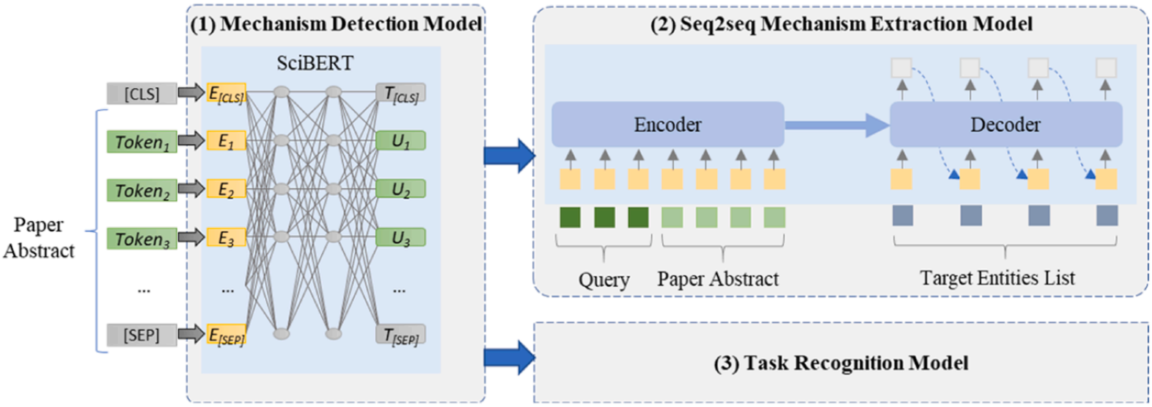


Fig. 4. The Metric-driven Mechanism Extraction Framework. (1): The mechanism detection model can filter papers that lack the metric-driven mechanism, which leverages SciBERT as the backbone; (2): We utilize the encoder-decoder architecture to extract a metric-driven mechanism. The seq2seq mechanism extraction model leverages BART as the backbone; (3): The task recognition model also leverages SciBERT, which is the same as the mechanism detection model, as the backbone. Note that we formalize the task entity extraction problem as a multi-label classification task to avoid entity normalization.

improve the robustness of the trained model, we augment the text data by randomly substituting or swapping words in the abstracts[9].

We employ the Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), an explainable AI (xAI) framework, to explain individual predictions. LIME first samples instances around an individual abstract text by adding a perturbation to the original text. Specifically, LIME randomly deletes words from the original text to perform perturbation, then classifies the generated samples using the trained model. Finally, the contribution of each word of the original text to the final model prediction result is generated by applying the LIME framework.

5.2. Mechanism Extraction

5.2.1. Task Formulation

Mechanism extraction is an information extraction task that includes named entity recognition and relation extraction. For a given abstract X_{abs}, our model should extract the metric-driven mechanism triples $\{(e_1, o_1, d_1), (e_2, o_2, d_2), \cdots, (e_i, o_i, d_i), \cdots, (e_n, o_n, d_n)\}$ from X_{abs}. The e_i, o_i, d_i correspond to the effect, operation, and direction in metric-driven mechanisms respectively, while n is the number of metric-driven mechanism triples.

The traditional methods formalize the NER as a sequence labeling or span classifying task. Given two entities and their related contexts, the RE is usually formalized as a multi-class classification task. The common paradigm of the NER and RE methods is to assign a label-specific classification layer based on pre-trained language models.

The metric-driven mechanism is distributed at the paragraph level, and many operation entities exceed ten. Moreover, the statement of metric-driven mechanisms is diverse. Therefore, we formalize the metric-driven mechanism as a text-to-text format, inspired by the T5 model (Raffel et al., 2020). T5 converts every language problem into a text-to-text format. Analogizing to multi-Turn question answering (X. Li et al., 2019), we cast the metric-driven mechanism extraction into a query-guided multi-turn text generation task.

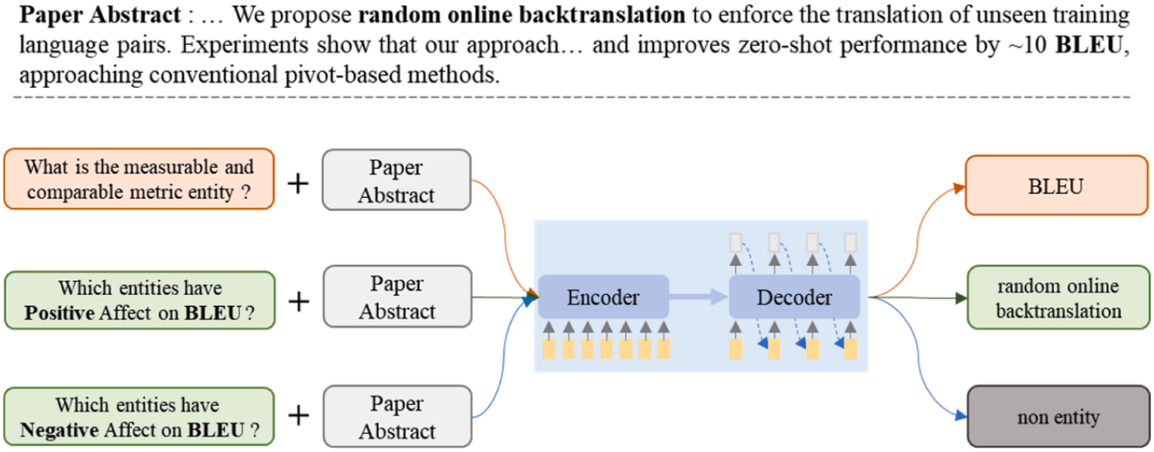
As shown in Fig. 5, for a given paper abstract, we extract the metric-driven mechanism by constructing a specific query for the *Effect* and *Operation*, which encodes vital information for the entity/relation class that we wish to identify. The query is concatenated with a paper's abstract to feed into the model. The model-generated text is the metric-driven mechanism's target entity.

5.2.2. The Query-guided Seq2seq Mechanism Extraction Model

To extract a metric-driven mechanism, we utilize the encoder-decoder architecture. Specifically, our mechanism extraction model employs BART (Lewis et al., 2020) as its backbone. To guide the model generation text, we adopt the encoder-decoder architecture with queries to generate the target entity in a metric-driven mechanism.

Encoder In our model, we design $Q = \{q_{effect}, q_{operation}\}$ as the model generating guide. Q encodes the semantic information on a specific entity/relation class. For a given X_{abs}, our model generates target e_i, o_i in multi-turns. The effect is the core of metric-driven mechanisms. Therefore, our model first concatenates the X_{abs} and q_{effect} as encoder input X_{q+abs} to obtain the effect entities $\{e_1, e_2, \cdots, e_m\}$. Then, we construct the query $q_{operation} = \{q_+, q_-, q_\sim\}$ for the *Operation* extraction. As shown in Fig. 5, the query for *Operation* incorporates the information about the *Effect* and *Effect* change direction. q_+, q_-, and $q_\sim$ correspond to an *Operation* that has positive, negative, and other effects on a specific *Effect* entity.

[9] We use the nlpaug (https://nlpaug.readthedocs.io) to augment the text data.



Paper Abstract : … We propose **random online backtranslation** to enforce the translation of unseen training language pairs. Experiments show that our approach… and improves zero-shot performance by ~10 **BLEU**, approaching conventional pivot-based methods.

What is the measurable and comparable metric entity ?
+
Paper Abstract
Which entities have **Positive** Affect on **BLEU** ?
+
Paper Abstract
Which entities have **Negative** Affect on **BLEU** ?
+
Paper Abstract
Encoder
Decoder
BLEU
random online backtranslation
non entity

Fig. 5. We jointly extract the entities (Operation and Effect) and relation in metric-driven mechanism by constructing different queries to guide the seq2seq mechanism extraction model to generate target entities. The "non entity" refers to the fact that no related operation entity has a negative effect on BLEU. If multiple target entities exist, these entities will be concatenated by a specific separator.

The encoder in the encoder-decoder architecture is used to encode the X_{q+abs} into the hidden representation space as vector H_{en}.

$$H_{en} = Encoder(X_{q+abs}) \tag{4}$$

where, $H_{en} \in \mathscr{R}^{n \times d}$ and d is the hidden state dimension.

Decoder In our model, the target sequence Y refers to the entities in the metric-driven mechanism. For a given query, if the number of target entities exceeds one, such as a paper's abstract that has more than one effect entity, we use "<>" as a separator between the target entities. The decoder uses the encoder outputs H_{en} and the previous decoder outputs $y_1, y_2, y_3, \cdot, y_{t-1}$ as the inputs and outputs the hidden state h_t for y_t.

$$h_t = Decoder\left(H_{en}; \widetilde{y}_{i=1}^{t-1}\right) \tag{5}$$

where $h_t \in \mathscr{R}^d$, $\widetilde{y}_{i=1}^{t-1}$ is the decoder outputs before t.

5.3. The Task Recognition Model

Traditionally, the task entity extraction problem is formalized as a sequence-labeling task. However, the sequence-labeling methods extract the entity without normalization. For example, the extracted entities "NER" and "named entity recognition" are considered different entities. To avoid uncontrollable extracted task entities and entity normalization, we formalize the task entity extraction problem as a multi-label classification task.

The architecture of the task recognition model is similar to the mechanism detection model, which consists of a text encoder and a classification layer. Here, we use SciBERT as a text encoder, which converts the input X_{abs} to an embedding h. Then, we use the fully connected layer to perform dimension mapping and the softmax to obtain the probability distribution for all 2,328 task categories.

$$p = Softmax(W \cdot h + b) \tag{6}$$

where p is a 2328-dimensional vector that denotes the probability distribution on task entities and **W** and **b** denote the weight and bias in the fully connected layer, respectively.

6. Model Evaluation

6.1. Experimental Setup

We ran experiments using the implementations of pre-trained models from Huggingface transformers[10]. Specifically, the BART model is distilbart-cnn-12-6, and the T5 model is T5-base. We use the PyTorch code for SpERT[11]. For text preprocessing, we use the Spacy. The batch size is 8. The number of training epochs is 40. The generation max length of the Seq2seq Mechanism Extraction Model is 20. The learning rate is 2 $\times 10^{-5}$. All of the methods were run on a server configured with RTX 3090Ti, 32CPU, and 32G memory.

[10] https://huggingface.co.

[11] https://github.com/lavis-nlp/spert.

6.2. Evaluation of the BERT Based Mechanism Detection Model

In the training stage, we built the Abs+Sent dataset by adding sentence granularity text to the original paragraph granularity training dataset. Specifically, if the sentence contained an effect entity, it was labeled 1. Additionally, we built the Aug-Abs+Sent dataset by randomly substituting or swapping words in the abstracts of Abs+Sent dataset.

As shown in Fig. 6, we find that incorporating sentence granularity text (Abs+Sent in Fig. 6) improves the mechanism detection performance compared with using the papers' abstracts alone. Moreover, data augmentation also improved the performance of mechanism detection in recall. Our mechanism detection model achieves an 84.6 recall for papers abstracts that contain a metric-driven mechanism and a 72.2 F1-score for the test set, which has 136 papers' abstracts that contain a metric-driven mechanism and 199 that do not.

To verify the mechanism detection model, we adopted the LIME to interpret the mechanism detection model. Specifically, we employed LIME to explain the mechanism detection model predictions for an abstract. Based on the LIME framework, we find that cue verbs and metrics entities significantly help to judge whether or not an abstract contains a metric-driven mechanism. The cue verbs include "achieve", "increase", "improve", and "reduce". In Fig. 7, the model mainly uses "BLEU", "achieves", and "improve" to perform the mechanism detection.

6.3. Evaluation of the Query-guided Seq2seq Mechanism Extraction Model

We leverage external data, namely the entities data in the SCIERC dataset, to fine-tune the BART model for domain adaptation before the model training. SICERC is a collection of 500 scientific abstract annotated with scientific entities and relations. Benefiting from the encoder decoder architecture, we only need to construct a specific-query for the entity classes in SCIERC instead of changing the model structure for model domain adaptation fine-tuning. However, the SCIERC dataset is small, which resulted in the model that was fine-tuned on SCIERC achieving a lower performance than the model without fine-tuning for direction classification, as shown in Table 6. Moreover, we also augment the training text to improve the robustness and performance of our model.

We compared our mechanism extraction model with the BERTNER and SpERT. The BERTNER model is a general model in the scientific named entity recognition task, which is fine-tuned on the SciBERT. The SpERT (Eberts & Ulges, 2020) is a strong model for span-based joint entity and relation extraction in a sequence-labeling style.

Qualitative analysis. Our mechanism extraction model achieves high fidelity. For the generated metric-driven mechanism entities, 94% of the effect entities and 77% of the operation entities can be found in the original abstract text, respectively. The huge number and diversity of operation entities cause lower fidelity compared to the effect entities. Moreover, there are two reasons why a few of the generated entities are not drawn from the original text.

1 Our model paraphrases the target entity. For example, the "structured" in "structured neural parser" is substituted by "structural" in our generated entity based on our seq2seq mechanism extraction model.
2 Our model does not simply copy the original text to overfit the original entities but generates target entities through reasoning. For example, our model generates "accuracy" from "we show that scaling to large topic spaces results in much more accurate models."

Quantitative analysis. From the perspective of text generation, our model achieves 53 on the rouge-1 score and 37 on the rouge-2 score. We used the relax match to compute the generated entities. If the word overlap rate between the reference entity and the generated entity exceeds 0.9, the generated entity will be considered a correct entity. As shown in Table 6, our BART-based seq2seq mechanism extraction model achieves a 63.6 F1-score on Operation and Effect entity recognition and a 49.4 F1-score on Direction recognition. Metric entities are tagged in SCIERC. Therefore, the fine-tuned BART-based seq2seq model achieves an 86.5 F1-score on effect entities. Compared with the span-based NER methods (Fu et al., 2021; Z. Jiang et al., 2020) which are restricted by the max span length, our model can generate long entities without any length restriction.

Additionally, we change the backbone to the T5 model (Raffel et al., 2020), which is pre-trained on a multi-task mixture of unsupervised and supervised tasks. Each task is converted into a text-to-text format. The T5-based metric-driven mechanism extraction model achieves a 61.8 F1-score on Operation and Effect entity recognition and a 56.0 F1-score on Direction recognition, which outperforms the baseline model.

Our metric-driven mechanism extraction task requires the model to understand the input text and copy the target entity from the original text. The BART model uses a seq2seq architecture and is particularly effective in the case of text generation and comprehension tasks. The T5 model also uses the seq2seq architecture and converts each task into a text-to-text format. Compared with the BART-based model, the performance of the T5-based model is lower for entity recognition and higher for Direction relation recognition compared to the BART-based model. Based on an analysis of model output, we find that the text generated by the BART-based model is both more fluent and more coherent.

6.4. Evaluation of the BERT Based Task Extraction Model

Our BERT-based task extraction model achieves an 89 F1-score, 93 on precision, and 85 on recall. The paper task distribution is unbalanced. We divided the task into three classes: high-, middle-, and low-frequency tasks. According to the frequency of the tasks in the total dataset, the high-frequency tasks are the top 25% of the tasks, the low-frequency tasks the bottom 50%, and the rest are middle-frequency tasks. Furthermore, we divide the papers included within the test dataset into three categories. As shown in Table 7,

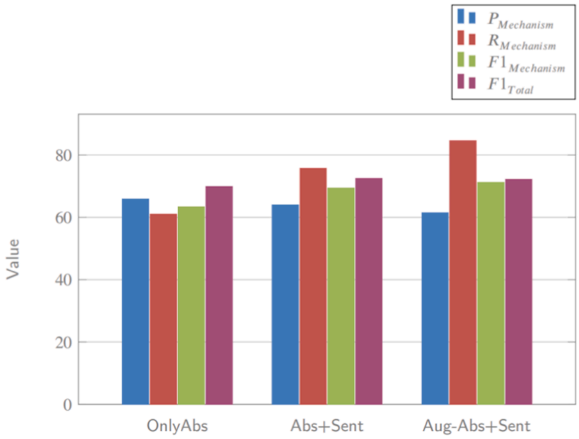
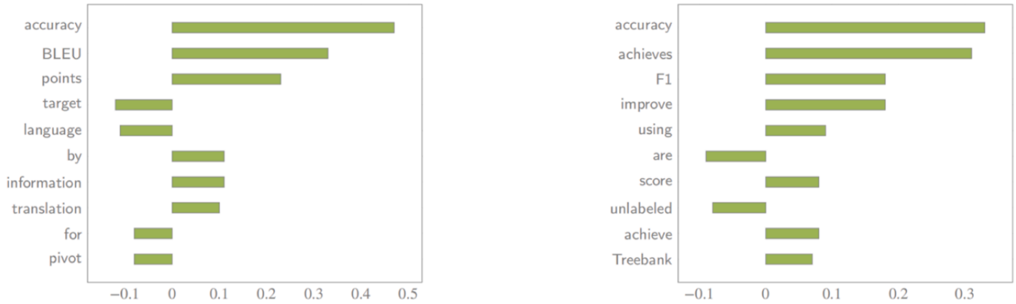


Fig. 6. Result of the BERT-based mechanism detection model. We primarily focused on the test results for papers' abstracts containing metric-driven mechanisms. $\boldsymbol{P_{Mechanism}}$, $\boldsymbol{R_{Mechanism}}$, $\boldsymbol{F1_{Mechanism}}$ correspond to the precision, recall, and F1-score for papers abstracts containing a metric-driven mechanism respectively. $\boldsymbol{F1_{Total}}$ refers to the F1-score for all of the papers' abstracts.



(a) Abstract 1: ··· In this paper, we propose a novel approach to remember the pivot phrases in the triangulation stage, and use a pivot language model as an additional information source at translation time. Experimental results on the Europarl corpus showed gains of 0.4-1.2 BLEU points in all tested combinations of languages.

(b) Abstract 1: ··· By incorporating a mixture of labeled and unlabeled data, we are able to improve relation classification accuracy, reduce the need for annotated data, ··· . We achieve this using a latent variable model that is trained in a reduced dimensionality subspace using spectral methods. Our approach achieves an F1-score of 0.485 on ···.

Fig. 7. Examples of the mechanism detection model based on the LIME framework. The x-axis refers to the word's contribution to the prediction result, where the positive and negative values correspond to the probability that the abstract text does or does not contain a metric-driven mechanism, respectively. a) and b) are the abstracts in natural language processing.

Table 6
Evaluation result of the metric-driven mechanism extraction model.

	Effect			Operation			Total Entities			Direction		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERTNER	63.7	76.6	69.6	38.1	60.0	46.6	50.9	68.3	58.1	-	-	-
SpERT	69.2	67.0	68.1	56.7	50.7	53.5	63.0	58.8	60.8	35.2	35.1	35.2
Seq2Seq_{NoFinetune}	79.0	81.6	80.3	49.4	49.4	49.4	62.0	63.1	62.6	49.4	49.4	**49.4**
Seq2Seq_{Finetune}	84.7	88.4	86.5	47.2	46.1	46.6	63.2	64.1	**63.6**	47.2	46.1	46.6

our model outperforms on papers with high-frequency tasks, but only achieves a 50 F1-score for papers with low-frequency tasks.

7. The NLP Metric-driven Mechanism KG

7.1. The Construction of the NLP Mechanism KG

Task entities refer to the research problem in NLP scientific papers. We further extend the metric-driven mechanism from

Table 7
Evaluation result of the BERT-based task extraction model.

	P	R	F1	# of Papers
High	94.5	86.8	90.5	4,810
Middle	86.0	65.2	74.2	530
Low	69.2	39.1	50.0	144
Total	93.1	85.7	89.2	5,000

(Operation, Effect, Direction) to (Operation, Effect, Direction, Task), which is called the n-ary mechanism relation. Using the PWC hierarchical task taxonomy, our NLP metric-driven mechanism KG supports the automatic semantic extension of tasks. For example, our NLP Mechanism KG can yield metric-driven mechanisms for *Paraphrase Generation, News Generation*, and *Paper generation* when a user sends a query about *Text Generation*. We find that the text generated by the BART-based model is both more fluent and more coherent compared to the T5-based model. Therefore, we use the BART-based model to extract metric-driven mechanisms.

Fig. 8a shows that the proposed NLP metric-driven mechanism KG contains three entity classes: Tasks, Operations, and Effects. Moreover, our knowledge graph contains four relation classes (“Positive,” “Negative,” “Other,” and “evaluated By”). “Positive,” “Negative,” and “Other” describe the influence direction between the operation entity and the effect entity, while “evaluated By” denotes the relationship between the task and effect entities. Based on the papers’ abstracts, that were downloaded from Semantic Scholar Academic Graph[12], we build a knowledge graph of the metric-driven mechanisms in the NLP domain (MKG_{NLP}). There are a few duplicated metric-driven mechanisms, i.e., two mechanisms have the same operation and effect entities, but the relations between the entities differ. We propose a relation priority rule (*negative* > *positive* > *other*) to eliminate the duplicated metric-driven mechanisms. Finally, the MKG_{NLP} has 43K n-ary mechanism relations.

7.2. Analysis of the NLP Mechanism KG

We selected 7,908 abstracts from papers that were published at ACL conferences between 2000 and 2021 to analyze the metric-driven mechanisms distribution. As shown in Fig. 8b, 40% of the papers’ abstracts contain a metric-driven mechanism. This proportion of papers shows an upward trend, and 6,825 n-ary mechanisms were extracted from the papers’ abstracts. To check the quality of the extracted metric-driven mechanisms, we manually verified 70 metric-driven mechanisms from 50 ACL 2015 papers. The accuracy of the extracted metric-driven mechanisms is 81.4%.

In our KG, after manually merging the same entities for entity disambiguation, the top ten effect entities were found to be “performance,” “accuracy,” “BLEU,” “precision,” “translation quality,” “robustness,” “recall,” “perplexity,” “error,” and “speed,” in descending order of occurrence. Moreover, a positive change direction of an Effect entity under the Operation entity accounts for the majority, accounting for 53% of the total, while a negative change direction accounts for 8%. There are 21,014 operation entities, so these far outnumbered the effect entities. For the operation entities, we use the pre-trained language model to obtain the embedding of operation entities, which were then clustered based on their embeddings. In Fig. 9, we visualize the results of the operation entities clustering, and the representative operation entities in Fig. 9 are shown in Table 8.

7.3. Applications of MKG_{NLP}

The MKG_{NLP} enables applications to retrieve metric-driven mechanisms in NLP to obtain the procedural scientific information on how to optimize the quantitative metrics of a specific task. For example, a user can search all papers that contain a mechanism related to the question: *how to improve the BLEU score for the machine translation task.* The Google Scholar search result items are unstructured snippets from scientific abstracts as shown in Fig. 10a. We design a prototype search engine for retrieving the metric-driven mechanism to illustrate the application value of MKG_{NLP}. Our metric-driven mechanism search engine can provide user structured, fine-grained metric-optimization information, which improves the efficiency of searching as shown in Fig. 10.

The MKG_{NLP} is acted as the back-end data server, which provides the metric-driven mechanism. The front-end interface is shown in Fig. 10b, in which users can send the query by filling the slots (effect, direction and task). To obtain target metric-driven mechanisms, we employ the SciBERT to obtain the embeddings of effect and task entities in MKG_{NLP}, then compute the cosine similarity score to find potentially relevant n-ary mechanism relations. Finally, we re-rank the retrieved relations according to the papers’ year of publication.

8. Implications and Conclusion

8.1. Implications

This study has the following theoretical implications. First, we propose a metric-driven mechanism schema in the form of (Operation, Effect, Direction) to represent the metric-optimization information. It bridges the gap between the current scientific

[12] https://www.semanticscholar.org/product/api. We query the API using the paper id in aclanthology.org, such as K17-1003. Aclanthology holds not only papers published at the ACL conference but also other computational linguistics conferences and artificial intelligence conferences.

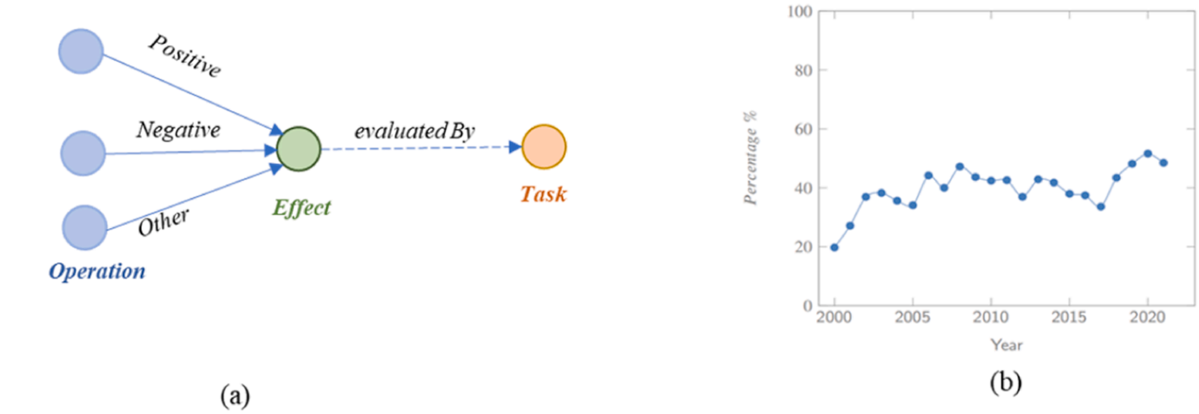


Fig. 8. Figure a is the schematic of entities and relations in the NLP Mechanism KG. Figure b is the chart about the percentage of papers with metric-driven mechanisms in ACL conference per year.

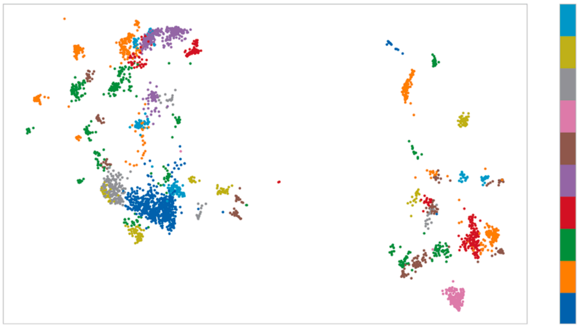


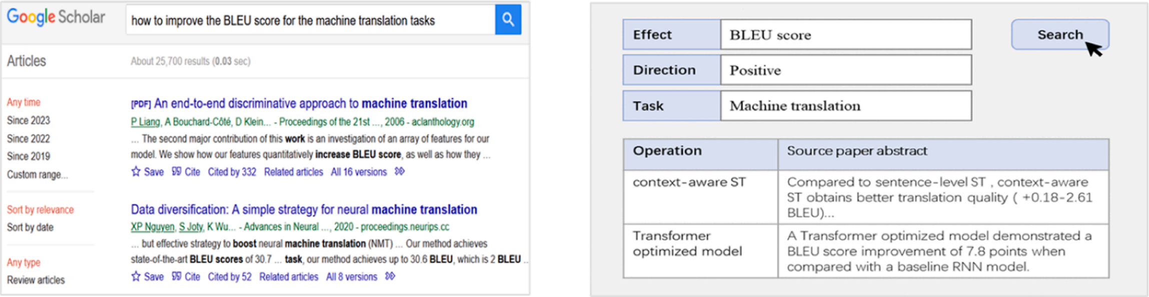
Fig. 9. Visualization of the operation entities. The number of clusters is set at 10.

Table 8
The representative operation entities for every topic.

Topic Id	Representative Entities
1	independence of applications, long intensity of the breath segments, use of overlapping phrases
2	Birectional Recurrent Neural Networks with LSTM cells, hierarchical recurrent neural network model, end-to-end recurrent neural network (RNN) models
3	exposure bias, human bias of comment quality, meta data
4	sequential classifier, unsupervised framework, global classifiers
5	a probability-threshold method, a tree-based model, decoder-only architecture
6	using the TF-IDF-weighted character n-gram model, tf-idf vectors, TF-IDF character n-grams
7	incorporate features extracted from learned nominals and their contexts, add multilingual links between speech segments in different languages, leveraging multilingualism and abundant monolingual corpora
8	back-transliteration, machine-generated questions, incorporate both document structure and PICO query formulation
9	TALP-UPC system, UPM system, UNT HiLT+Ling system
10	combining statistical approaches, combination of two separate encoder,Combined method

information extraction schema and the information need in applied artificial intelligence, i.e., from answering factual questions beginning with the word “what” to answering procedural questions beginning with the word “how”. Second, different from previous scientific information extraction works which extract entities and relations in a sequence-labeling manner, our proposed framework extracts metric-driven mechanisms in the form of multi-turn text generation.

In terms of practical implications, this study has potential for creating applications to assist applied AI scientists to solve specific problems. Since previous information extraction schemas and methods cannot directly provide users with procedural information, extracting the procedural scientific information contained in scientific publications, particularly the metric-optimization information, can improve the efficiency of searching, reading, and using. Moreover, the mechanism extraction model presented in this paper is a general model, which can be easily transferred to other domains with the accordingly adjusted mechanism schema.



how to improve the BLEU score for the machine translation tasks

About 25,700 results (**0.03** sec)

[PDF] An end-to-end discriminative approach to **machine translation**
P Liang, A Bouchard-Côté, D Klein... - Proceedings of the 21st ..., 2006 - aclanthology.org
... The second major contribution of this **work** is an investigation of an array of features for our model. We show how our features quantitatively **increase BLEU score**, as well as how they ...
Cited by 332 All 16 versions

Data diversification: A simple strategy for neural **machine translation**
XP Nguyen, S Joty, K Wu... - Advances in Neural ..., 2020 - proceedings.neurips.cc
... but effective strategy to **boost** neural **machine translation** (NMT) ... Our method achieves state-of-the-art **BLEU scores** of 30.7 ... **task**, our method achieves up to 30.6 **BLEU**, which is 2 **BLEU** ...
Cited by 52 All 8 versions

Effect	BLEU score
Direction	Positive
Task	Machine translation

Operation	Source paper abstract
context-aware ST	Compared to sentence-level ST , context-aware ST obtains better translation quality (+0.18-2.61 BLEU)...
Transformer optimized model	A Transformer optimized model demonstrated a BLEU score improvement of 7.8 points when compared with a baseline RNN model.

(a) The search result of google scholar about “how to improve the BLEU score for the machine translation tasks.”

(b) The prototype for the metric-driven mechanism search engine.

Fig. 10. Comparison between our metric-driven mechanism search engine and Google Scholar.

8.2. Conclusion

In conclusion, we introduce a coarse-grained representation schema to express metric-driven mechanisms in the fields of AI. Our schema focuses on procedural scientific information related to the metric-optimization. Furthermore, we construct a dataset based on the papers' abstracts in NLP domain for mechanism detection and metric-driven mechanism extraction. We propose a framework based on the pre-trained model to extract the metric-driven mechanism and paper task. Considering the distribution and statement patterns of metric-driven mechanisms, we formalize the metric-driven mechanism triple as a query-guided multi-turn text generation task. Based on the proposed framework, a knowledge graph of metric-driven mechanisms in NLP (MKG_{NLP}) is constructed. Human evaluation shows that the extracted metric-driven mechanisms have an 81.4% accuracy. Moreover, there are 43K n-ary mechanism relations in the form of (Operation, Effect, Direction, Task) in our MKG_{NLP}. Finally, the metric-driven mechanism search engine shows the advantage of supporting applied AI scientists to solve the domain-specific problem.

There exist many metric-driven mechanisms in the main sections of papers, such as the results and discussion sections. In future, we will explore metric-driven mechanism extraction from paper's main sections. A high-quality annotated NLP dataset in science is scarce, so we will explore utilizing self-supervised tasks to fine-tune the pre-trained model. Furthermore, we will also explore more applications based on the metric-driven mechanism knowledge graph.

CRediT authorship contribution statement

Yongqiang Ma: Writing – original draft, Methodology, Data curation, Software. **Jiawei Liu:** Formal analysis, Validation. **Wei Lu:** Conceptualization, Supervision, Investigation. **Qikai Cheng:** Conceptualization, Supervision, Investigation, Validation.

Data availability

Data will be made available on request.

Acknowledgments

This research has been made possible through the support of the National Natural Science Foundation of China under Grants No. 72234005 and No. 72174157.

References

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 546–555). https://doi.org/10.18653/v1/S17-2091

Bechtel, W. (2010). The downs and ups of mechanistic research: Circadian rhythm research as an exemplar. *Erkenntnis, 73*(3), 313–328. https://doi.org/10.1007/s10670-010-9234-2

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3615–3620). https://doi.org/10.18653/v1/D19-1371

Chan, Y. S., & Roth, D. (2011). Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 551–560).

Chen, V.Z., Montano-Campos, F., & Zadrozny, W. (2020). Causal knowledge extraction from scholarly papers in social sciences. ArXiv Preprint, abs/2006.08904.

Chen, X., Zhang, N., Li, L., Xie, X., Deng, S., Tan, C., et al. (2022). Lightner: A lightweight generative framework with prompt-guided attention for low-resource NER. In *Proceedings of the 26th conference on computational natural language learning*.

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-based named entity recognition using BART. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 1835–1845. 10.18653/v1/2021.findings-acl.161.

Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., & Sack, H. (2020). Ai-kg: An automatically generated knowledge graph of artificial intelligence. *Proceedings of the international semantic web conference*, 127–143. https://doi.org/10.1007/978-3-030-62466-8_9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and short papers)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M. Y., Auer, S., & Ewerth, R (2020). The STEM-ECR dataset: Grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2192–2203).

Eberts, M., & Ulges, A (2020). Span-based joint entity and relation extraction with transformer pre-training. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, & J. Lang (Eds.)*, 325. Proceedings of the ECAI 2020-24th European conference on artificial intelligence* (pp. 2006–2013). IOS Press. https://doi.org/10.3233/FAIA200321, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29-September 8, 2020-Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)Vol.

Fan, T., & Wang, H. (2022). Research of Chinese intangible cultural heritage knowledge graph construction and attribute value extraction with graph attention network. *Information Processing & Management, 59*(1), Article 102753. https://doi.org/10.1016/j.ipm.2021.102753

Fu, J., Huang, X., & Liu, P. (2021). SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)* (pp. 7183–7195). https://doi.org/10.18653/v1/2021.acl-long.558

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018). SemEval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 679–688). https://doi.org/10.18653/v1/S18-1111

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*(1). https://doi.org/10.1007/BF00172853

Hope, T., Amini, A., Wadden, D., van Zuylen, M., Parasa, S., Horvitz, E., et al. (2021). Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 4489–4503). https://doi.org/10.18653/v1/2021.naacl-main.355

Hope, T., Chan, J., Kittur, A., & Shahaf, D. (2017). Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 235–243). https://doi.org/10.1145/3097983.3098038. Halifax, NS, Canada, August 13 - 17, 2017.

Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2021). TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 707–714). https://doi.org/10.18653/v1/2021.eacl-main.59

Huo, C., Ma, S., & Liu, X. (2022). Hotness prediction of scientific topics based on a bibliographic knowledge graph. *Information Processing & Management, 59*(4), Article 102980. https://doi.org/10.1016/j.ipm.2022.102980

Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7506–7516). https://doi.org/10.18653/v1/2020.acl-main.670

Ji, D., Tao, P., Fei, H., & Ren, Y. (2020). An end-to-end joint model for evidence information extraction from court record document. *Information Processing & Management, 57*(6), Article 102305. https://doi.org/10.1016/j.ipm.2020.102305

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems, 33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Jiang, Y., Huang, Y., Xia, Y., Li, P., & Lu, W. (2021). Recognition of lexical functions in academic texts: Application in automatic keyword extraction. *Journal of the China Society for Scientific and Technical Information, 40*(2), 152. https://doi.org/10.3772/j.issn.1000-0135.2021.02.005

Jiang, Z., Xu, W., Araki, J., & Neubig, G. (2020). Generalizing natural language analysis through span-relation representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2120–2133). https://doi.org/10.18653/v1/2020.acl-main.192

Kabongo, S., D'Souza, J., & Auer, S (2021). Automated mining of leaderboards for empirical AI research. In H.-R. Ke, C. S. Lee, & K. Sugiyama (Eds.), *Towards open and trustworthy digital societies* (pp. 453–470). Springer International Publishing.

Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T. (2010). SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 21–26).

Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. In J. Vanschoren, & S. Yeung (Eds.)*, 1. Proceedings of the neural information processing systems track on datasets and benchmarks.* Vol.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing, 470*, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). https://doi.org/10.18653/v1/2020.acl-main.703

Li, P., Liu, Q., Cheng, Q., & Lu, W. (2021). Data set entity recognition based on distant supervision. *The Electronic Library, 39*(2).

Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., et al. (2019). Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1340–1350). https://doi.org/10.18653/v1/P19-1129

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: long papers)* (pp. 2124–2133). https://doi.org/10.18653/v1/P16-1200

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. et al. (2019). Roberta: A robustly optimized bert pretraining approach. ArXiv Preprint, abs/1907.11692.

Lu, W., Li, P., Zhang, G., & Cheng, Q. (2020). Recognition of lexical functions in academic texts: Automatic classification of keywords based on BERT vectorization. *Journal of the China Society for Scientific and Technical Information, 39*(12), 1320. https://doi.org/10.3772/j.issn.1000-0135.2020.12.008

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 3219–3232). https://doi.org/10.18653/v1/D18-1360

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Martínez-Plumed, F., Barredo, P., hÉigeartaigh, S.Ó., & Hernández-Orallo, J. (2021). Research community dynamics behind popular AI benchmarks. *Nature Machine Intelligence, 3*(7), 581–589. https://doi.org/10.1038/s42256-021-00339-6

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence, 171*(18), 1174–1182. https://doi.org/10.1016/j.artint.2007.10.009

Mondal, I., Hou, Y., & Jochim, C. (2021). End-to-End construction of NLP knowledge graph. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 1885–1895. 10.18653/v1/2021.findings-acl.165.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). San Francisco, CA, USA: ACM. https://doi.org/10.1145/2939672.2939778. August 13-17, 2016.

Röhl, J. (2012). Mechanisms in biomedical ontology. *Journal of Biomedical Semantics, 3*(2), S9. https://doi.org/10.1186/2041-1480-3-S2-S9

Schlangen, D. (2021). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 2: short papers)* (pp. 670–674). https://doi.org/10.18653/v1/2021.acl-short.85

Steel, D. (2007). *Across the boundaries: Extrapolation in biology and social science.* Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). https://doi.org/10.5555/3295222.3295349

Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5784–5789). https://doi.org/10.18653/v1/D19-1585

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., et al. (2020). CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020*.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various NER subtasks. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (Volume 1: long papers)* (pp. 5808–5822). https://doi.org/10.18653/v1/2021.acl-long.451

Yang, F., Moss, L. G., & Phillips, G. N. (1996). The molecular structure of green fluorescent protein. *Nature Biotechnology, 14*(10), 1246–1251. https://doi.org/10.1038/nbt1096-1246

Zheng, A., Zhao, H., Luo, Z., Feng, C., Liu, X., & Ye, Y. (2021). Improving on-line scientific resource profiling by exploiting resource citation information in the literature. *Information Processing & Management, 58*(5), Article 102638. https://doi.org/10.1016/j.ipm.2021.102638

Zhong, Z., & Chen, D. (2021). A frustratingly easy approach for entity and relation extraction. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 50–61). https://doi.org/10.18653/v1/2021.naacl-main.5