CEUR-WS.org/Vol-3210/paper1.pdf

Beyond Tasks, Methods, and Metrics: Extracting Metrics-driven Mechanism From the Abstracts of AI Articles

Yongqiang Ma*, Jiawei Liu*, Wei Lu and Qikai Cheng**

Wuhan University, Wuhan, Hubei Province, P.R.China. 430072

Abstract

Generally speaking, a scientific paper presents the result of a specific research area and provides a solution to the research question. With the exponential expansion in the number of scientific publications, a large amount of valuable information is submerged. Although existing information extraction methods can extract entities and relations, they are unable to directly provide readers with the mechanism that reveals the path to solve the problem. Inspired by the biomedical research of medical mechanism, in this paper, we propose a novel knowledge schema, i.e., metrics-driven mechanism knowledge schema (***Operation, Effect, Direction***), which depict the knowledge about *"How to optimize the quantitative and qualitative metrics of a specific task?"*. Furthermore, we select the natural language processing domain for practice, which is a representative branch of Artificial Intelligence (AI). Specifically, we construct a mechanism sentence extraction dataset and a mechanism triple extraction dataset using abstract data from ACL papers based on the proposed schema. Then we propose a metrics-driven mechanism knowledge extraction pipeline based on the pre-trained model. Finally, a knowledge base of metrics-driven mechanisms in the natural language processing (NLP) domain, named NLPMKB, is constructed. The human evaluation results show that the extracted mechanism knowledge from NLPMKB is high-quality with 87.0% precision and 79.4% recall. Moreover, the experiments on the knowledge retrieval task demonstrate that the performance can be further improved with the support of the NLPMKB.

Keywords

Metrics-driven mechanism, Information extraction, Scientific papers mining

1. Introduction

> "Much of the practice of science can be understood in terms of the discovery and description of mechanisms." — Machamer et al. [1]

As a kind of knowledge, mechanism reveals how to manipulate things and help people understand the path to solve the problem[2, 1]. With the prosperous development of Artificial Intelligence (AI), the research field of AI is rapidly extending in multiple disciplines, and the corresponding research publications are growing rapidly. From the perspective of solving problems, AI research can be regarded as a work of discovering and describing the mechanisms between a specific problem and the corresponding solution [3]. In this paper, we explore the metrics-driven mechanism extraction from the abstract of AI articles.

Benchmarks [4] formalize a particular task through datasets and associated quantitative evaluation metrics. Improving the performance reflected by evaluation metrics on established benchmarks is an important way to increase the legitimacy of a research work [5]. Inspired by the definition of the benchmark-driven methodology [6], we define the metrics-driven research pattern in AI as the pattern that focuses on optimizing the performance of a particular task reflected by the quantitative and qualitative metrics. Therefore, the metrics-driven mechanism in our paper can be regarded as knowledge about how to optimize the quantitative and qualitative metrics of a specific task.

Generally speaking, readers have specific questions when reading scientific publications [7]. For instance, *Can the mechanism knowledge discovered in this paper solve my problem?* is the general question that arises when readers read scientific papers to solve problems using AI knowledge. In the past 20 years, the number of scientific publications in the AI domain has grown twelvefold [8]. As a result, a large number of valuable mechanism knowledge has unfortunately been submerged in the "information flood". Since a large amount of information is encoded in a paper in the form of text, it is difficult for people to obtain mechanism knowledge when facing "information overload". Therefore, it is evident that extracting the information contained in scientific publications can improve the efficiency of searching and reading when facing a specific question, as shown in Figure 1.

Research in natural language processing (NLP) has provided great convenience in terms of extracting fine-

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online

*Both authors contributed equally to this research.

**Corresponding author.

EMAIL: mayongqiang@whu.edu.cn (Yongqiang Ma); laujames2017@whu.edu.cn (Jiawei Liu); weilu@whu.edu.cn (Wei Lu); chengqikai0806@163.com (Qikai Cheng)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

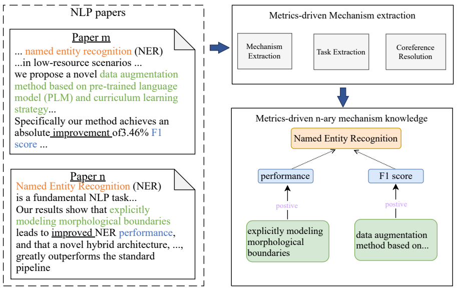

NLP papers
Paper m
... named entity recognition (NER)
...in low-resource scenarios ...
we propose a novel data augmentation method based on pre-trained language model (PLM) and curriculum learning strategy...
Specifically our method achieves an absolute improvement of3.46% F1 score ...
Paper n
Named Entity Recognition (NER) is a fundamental NLP task...
Our results show that explicitly modeling morphological boundaries leads to improved NER performance, and that a novel hybrid architecture, ..., greatly outperforms the standard pipeline
Metrics-driven Mechanism extraction
Mechanism Extraction
Task Extraction
Coreference Resolution
Metrics-driven n-ary mechanism knowledge
Named Entity Recognition
performance
F1 score
postive
postive
explicitly modeling morphological boundaries
data augmentation method based on...

Figure 1: Our model primarily focuses on extracting the metrics-driven mechanism knowledge in the abstract regarding the operations related to performance improvement and optimization. The orange block is the task entity, the green blocks are the operation entities, and the blue blocks are the effect entities. The arrows between the operation entity and the effect entity indicate a change in direction by the effect entity under the influence of the operation entity.

grained entities and relations in scientific publications, including task entity, method entity, dataset entity, and metric entity identification [9, 10, 11, 12, 13]; chemical entity recognition [14, 15]; and biomedical named entity recognition [16]. Most of the current works extract descriptive scientific information instead of procedural scientific information. For instance, descriptive scientific information includes the problems addressed in scientific publications, the domain of a research question, and the method used to address a problem [17]. As a kind of procedural scientific information, a discovered mechanism oriented toward improving quantitative metrics in AI is neglected in scientific information extraction.

Inspired by Chen et al. [18] and Hope et al. [19], we construct a metrics-driven mechanism knowledge representation schema to express the key procedural scientific information in AI. In our scheme, the metrics-driven mechanism knowledge is represented as a triple $(\boldsymbol{Operation}, \boldsymbol{Effect}, \boldsymbol{Direction})$. *Operation* is the entity that refers to the innovative model, method, or approach proposed in a paper. *Effect* is the entity that refers to the metrics evaluating operation's effectiveness or performance. *Direction*, regarded as the relationship between the operation entity and the effect entity, refers to the change in direction by the effect entity under the influence of the operation entity. Based on the trade-off between scalability and expression capabilities, we preliminarily divide the *Direction* into three categories: positive effect, negative effect, and other. The coarse-grained metrics-driven mechanism knowledge representation schema achieves a balance between domain adaptability and universality, which can be applied not only in AI but also in biology and chemistry.

This paper chooses the NLP domain for practice since it is a representative and prosperous branch of AI. First, we construct an annotated mechanism knowledge extraction dataset based on the abstracts of ACL papers[1]. Then, we propose a model that utilizes the pre-trained knowledge to extract metrics-driven mechanism triples. Finally, we construct a metrics-driven mechanism knowledge base in NLP, named NLPMKB, to further improve the performance of knowledge retrieval.

In summary, our primary contributions are as follows:

- We propose a coarse-grained metrics-driven mechanism knowledge representation schema. Moreover, based on the proposed schema, an annotated dataset is constructed in the field of NLP with 1,486 mechanism triples.
- Utilizing the annotated dataset, we train an information extraction (IE) model. The experimental results show that the BERT-based mechanism sentence extraction model can achieve an 89 F1 score, and the mechanism triple extraction model based on SpERT achieves 78.7 and 59.8 F1 score on the mechanism entity recognition task and the relation extraction task, respectively.
- Based on the trained model, a large number of publications in ACL are extracted to construct a metrics-driven mechanism knowledge base (KB) in the NLP domain. The human evaluation results show that our metrics-driven mechanism KB has high quality and utility. Our search engine achieves 20- and 12-point improvements on P@3 and P@5 in the metrics-driven mechanism knowledge retrieval task.

2. Related Work

2.1. Mechanism Knowledge in Science

Mechanisms are involved in the research of many disciplines. In biology, biochemists and molecular biologists pursue explanations of genes, proteins, and the molecules that influence biochemical reactions in the context of mechanistic explanations [20, 21, 22]. In chemistry, researchers regard chemical reactions as a mechanism.

According to the Oxford dictionary, a mechanism is *"a natural system or type of behavior that performs a particular function"*[2]. In the philosophy of science, there is a great deal of discussion about the formal definition of mechanism. For example, Machamer et al. [1] defined mechanisms as organized entities and activities that produce regular changes from start or set-up to finish or

[1] https://aclanthology.org

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

termination conditions. Glennan [2] defined the mechanism as a complex system that produces behavior through the interaction of several parts according to direct causal laws.

2.2. Information Extraction

Information extraction (IE) refers to the extraction of structured information from unstructured or semi-structured texts. In general, the problem of IE is composed of Named Entity Recognition (NER) and Relation Extraction (RE) tasks. There are two types of approaches for IE: the pipeline-based approach and the end-to-end joint approach.

As for the pipeline-based approach, the model first recognizes entities using a NER method and then extracts the relations between recognized entities with an RE method [23, 24, 25]. The strength of the pipeline-based approach is its flexibility when integrating different data sources and methods. However, its weakness is the error propagation problem between the individual NER step and RE step.

In terms of the end-to-end joint approach, the model jointly extracts entities and relations using the shared layer or shared parameters between the NER task and the RE task; such models include DygIE++ [26] and SpERT [27]). Moreover, Yan et al. [28] employed an encoder-decoder framework based on BART [29] to extract entities in the text. Li et al. [30] designed an alternative strategy in which they cast the entity-relation extraction as a multi-turn question-answering problem.

2.3. Scientific Information Extraction

Information extraction from scientific literature allows researchers to gain key insights from scientific documents. There are differences in the types of entities and relations in different fields.

In AI, current scientific information extraction research primarily focuses on extracting keyphrases [31, 32], lexical functions of keyphrases [33], fine-grained scientific entities (e.g. *Task, Method, Dataset*, and *Metric*) [12, 11, 34, 10, 35], and relations[36, 37, 38]. In SemEval 2017 Task-10, Augenstein et al. [36] proposed the hyponym-of and synonym-of relations. In SemEval 2018 Task-7, Gábor et al. [37] proposed the *usage, result, part-whole*, and *compare* relations. Recently, Mondal et al. [38] proposed the *evaluated-On* and *evaluated-By* relations.

In short, current scientific information extraction research emphasizes descriptive information (e.g., task entities, method entities, dataset entities, and the relations between them), which primarily focuses on declarative information instead of procedural information in academic publications. Our work further extends the procedural information of mechanism knowledge between proposed operation entities and performance metric entities as they are oriented to the specific problem.

There are several studies related to mechanism knowledge extraction and representation in a specific domain. Hope et al. [39] proposed a weak structural representation that describes an idea in product descriptions regarding purpose (*what they are trying to achieve*) and mechanism (*how they achieve that purpose*). Chen et al. [18] identified the hypothesis sentences from scientific documents in business and management. Then, they extracted cause and effect entities in those hypothesis sentences. Hope et al. [19] built a COVID-19 mechanism relations knowledge base, which includes activities, functions, and influences relations extracted from CORD-19 papers. In summary, what current studies have in common is that they construct a very simple mechanism knowledge representation schema, which is an optimal solution considering trade-offs in terms of ease of extraction, scalability, and coverage.

3. Data and Task

3.1. Schema of Mechanism

In many scientific fields, a detailed description of the mechanism is required to deliver a satisfactory explanation [1]. *Mechanism*, a kind of knowledge, reveals how to manipulate things, promotes the development of science, and aids researchers in understanding and solving problems.

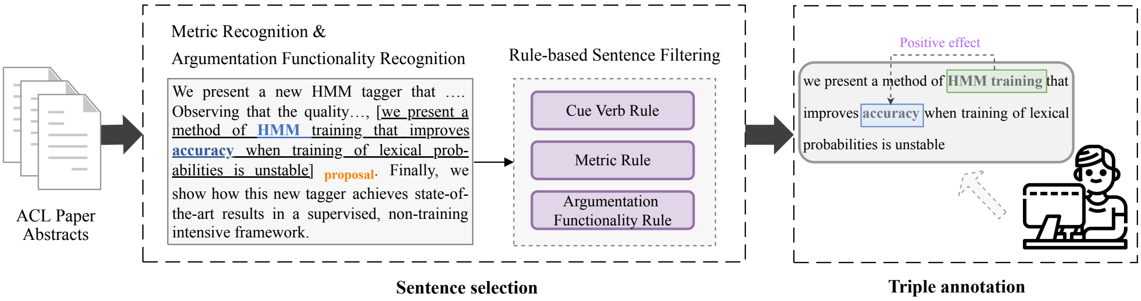
As shown in Table 1, mechanism knowledge exists in AI (e.g., natural language processing and computer vision), chemistry, biology, and other fields. Although the research fields are different, the common point core of mechanism knowledge is that it expresses the influential relationship between things or entities in a scheme. Whether the things or entities expressed in the mechanism are concrete (e.g., chemicals, cells, and plants) or abstract (e.g., theory and concept), we divide these things or entities into two types, *operation* and *effect*, based on the role in the mechanism.

We find that the mechanism knowledge in artificial intelligence research is primarily metric-driven, that is, it states the effect of the proposed methods and models on specific metrics as a key conclusion in the abstract of a scientific paper. The common expression forms primarily include the following two types based on the analysis and refinement of a large number of paper abstracts:

1. A direct description of the effect of the innovative model or method on the specific metric or aspect, such as: *model X improves (reduces/affects/achieves) metric M with specific change value (e.g., percentage).*

Table 1
The Mechanism knowledge in scientific research. The mechanism examples come from scientific abstracts in 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



We present a new HMM tagger that .... Observing that the quality..., [we present a method of HMM training that improves accuracy when training of lexical probabilities is unstable] proposal. Finally, we show how this new tagger achieves state-of-the-art results in a supervised, non-training intensive framework.

we present a method of HMM training that improves accuracy when training of lexical probabilities is unstable

Figure 2: Metrics-driven mechanism annotation process.

2. An indirect description of the effect of the innovative model or method on the specific metric or aspect by comparison, such as: *compared with the baseline, model X outperforms (or an adjective expressing the comparative degree) on M metric.*

In our schema, metrics-driven mechanism knowledge in the form of natural language can be abstracted as a triple $(\boldsymbol{Operation}, \boldsymbol{Effect}, \boldsymbol{Direction})$. This metric-driven mechanism triple represents an entity such as a model or a model proposed by the researcher, that corresponds to the “applying dictionary and web-based answer reranking together” in Figure 3. *Effect* in a metric-driven mechanism triple represents the metric entity, which corresponds to the “mean reciprocal rank score” in Figure 3. *Direction* expresses the relationship between the operation entity and the effect entity, which corresponds to the “increase” in Figure 4.

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

- *Positive effect*: 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.

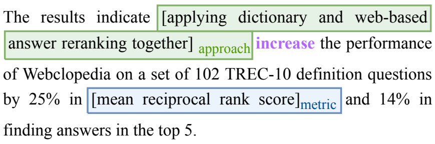

The results indicate [applying dictionary and web-based answer reranking together] approach increase the performance of Webclopedia on a set of 102 TREC-10 definition questions by 25% in [mean reciprocal rank score]metric and 14% in finding answers in the top 5.

Figure 3: Metrics-driven mechanism knowledge in a natural language form.

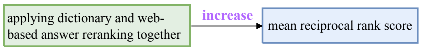

applying dictionary and web-based answer reranking together
increase
mean reciprocal rank score

Figure 4: Metrics-driven mechanism knowledge in a triplet form.

- *Negative effect*: the method/model proposed in the research article reduces the metric. Examples include using structural features to reduce the alignment error rate.
- *Other*: other than the above two relationships. For example, a external feature affect the metric but we did not know the effect direction.

3.2. Task Definition

To extract metrics-driven mechanism knowledge from the abstract text, we divide the target problem into two subtasks, i.e., selection-then-extraction corresponding to the dataset construction process:

Subtask 1 : mechanism knowledge sentence **selection**, which identifies sentences containing mechanism knowledge.

Subtask 2 : mechanism knowledge **extraction**, which extracts the mechanism knowledge triples from the recognized sentences.

3.3. Dataset

As shown in Figure 2, the construction of the mechanism knowledge extraction dataset primarily includes two steps: **sentence selection** and **triple annotation**.

The metrics-driven mechanism triples in our annotated dataset explicitly exist in a single sentence.Note that in our proposed dataset, the mechanism knowledge described across multiple sentences was excluded due to time and efficiency constraints. As shown in the following examples, the annotator did not consider the implicit mechanism knowledge existing between sentences in the process of annotation. In Examples 1 and 2, the effect entities (e.g., "accuracy" and "performance" in the examples) and operation entities (e.g., "using bilingual dictionary and transfer grammar" and "coarse-to-fine approach" in the examples) are separated from each other in different sentences.

Example 1 In Malayalam-Tamil pair, the divergence is more reported in lexical and structural level, that is been resolved by using **bilingual dictionary and transfer grammar**. The **accuracy** is increased to 65 percentage, which is promising.

Example 2 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.

3.3.1. Sentence selection

We find that most of the sentences containing the metrics-driven mechanism are distributes in the conclusion part, and it is intuitive that sentences containing the metrics-driven mechanism also contain the metric entities. Given a paper abstract, the annotator first needs to choose the target sentence that contains the metrics-driven mechanism knowledge. To improve annotation efficiency, three heuristic rules are proposed to detect possible target abstract sentences. Specifically, heuristic rules primarily consider three aspects: verbs, metric entities, and argumentation functionality types [40, 41, 42] in the sentence.

- **Cue Verb Rule**: verb words such as effect, influence, decrease, reduce, increase, and improve as well as their noun forms.
- **Metric Rule**: specific metric entities such as accuracy, F1 score, and BLEU as well as abstractive metric entities such as performance and quality.
- **Argumentation Functionality Rule**: the argument functionality of the sentence is the "proposal" or "outcome".

To apply these rules, we use SpERT[27] to recognize the metric entities in a sentence and trained a BERT-based argumentation functionality classifier based on the schema and dataset proposed by Accuosto and Saggion [42] in computational linguistics.

3.3.2. Triple annotation

Given a selected sentence containing metrics-driven knowledge, the annotator needs to label the entities in the metrics-driven mechanism's schema and then determine the relationship between entities based on the context.

We use brat[1] as the annotation tool for mechanism sentence recognition and mechanism knowledge tagging. The two annotators are graduate students with NLP backgrounds. For annotation disagreement on entity boundaries (e.g., "our model" vs. "model"), we choose the longer annotation (e.g., "our model"). The inter-annotator agreement score of our dataset is 0.952[2].

3.3.3. Annotated dataset analysis

Based on the annotated dataset, summaries of the statistics for the datasets for subtask 1 and subtask 2 are provided in Table 2 and Table 3. As shown in Table 2, the proportion of sentences containing mechanism knowledge is relatively low compared with non-mechanism

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

[2] The tool (https://github.com/kldtz/bratiaa) we adopted to calculate the F1 score of per document or label as the inter-annotator agreement score.

Table 2
Statistics of the dataset for subtask 1.

Statistics items	Number
Total # of Sentences	4,163
# of Mechanism Sentences	1,032
# of Non-mechanism Sentences	3,131
Avg # tokens	26

Table 3
Statistics of the dataset for subtask 2.

Statistics items	Number
Total # of Sentences	1,032
Avg # of Sentence Tokens	31
# of Entities	2,525
# of Operation Entities	1,214
Avg # of Operation Entity Tokens	3.02
# of Affect Entities	1,311
Avg # of Effect Entity Tokens	1.76
# of Relations	1,486
# of Pos Effect Relations	1,056
# of Neg Effect Relations	217
# of Affect Relations	213

sentences. We find that mechanism sentences are primarily distributed in the third to sixth sentence as shown in Figure 5. In addition, it can be found in Table 3 that the distribution of metric mechanism knowledge relations is also highly imbalanced, and positive effect relations account for the majority.

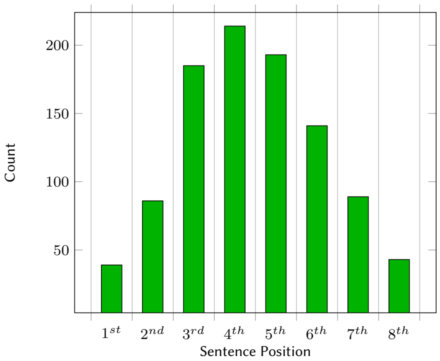


Figure 5: Position distribution of mechanism sentences in abstracts of papers.

4. KB Construction

We describe our approach, which is depicted in Figure 6, for extracting mechanisms from the abstracts of scientific papers. We first trained a metrics-driven mechanism knowledge extraction model based on the annotated dataset from a small collection of ACL papers (Section 3.3). Then, the trained model was applied to scientific papers in the natural language processing domain to build an NLP metrics-driven Mechanism Knowledge Base (**NLPMKB**), which supports and further improves the retrieval performance for metric-driven mechanisms. Finally, we built a metrics-driven mechanism knowledge search engine.

4.1. Extraction Pipeline for Mechanisms Knowledge

We propose a metrics-driven mechanism extraction pipeline that includes two steps: **mechanism sentence extraction** and **mechanism triple extraction**. Recently, pretrained language models, e.g., BERT[43], RoBERTa[44], and SciBERT[45], have promoted the performance of natural language understanding tasks ranging from text classification and named entity recognition to machine reading comprehension. SciBERT is a pretrained language model for scientific text, which leverages a large-scale scientific publications as a pretraining task dataset and advances downstream scientific NLP tasks. Therefore, our mechanism extraction pipeline uses SciBERT as a backbone for extracting the text's semantic information.

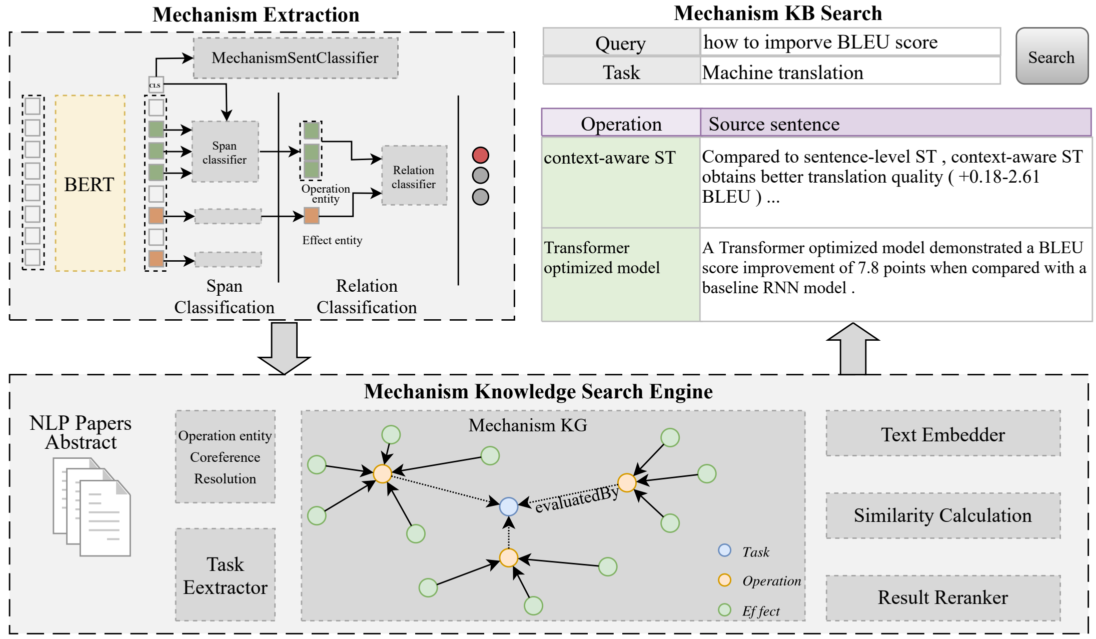
4.1.1. Mechanism Sentence Extraction

We formalize the mechanism sentence extraction as a binary text classification task. Given a sentence $sent$ in an abstract of a scientific paper, the model needs to identify whether $sent$ contains complete metric-driven mechanism knowledge. Our BERT-based mechanism sentence extraction model has two parts, i.e., text encoder and classification layer. We formalize the mechanism sentence extraction as a binary text classification task. Given a sentence in a scientific paper's abstract, the model needs to identify whether contains complete metrics-driven mechanism knowledge. Our BERT-based mechanism sentence extraction model has two parts: a text encoder and a classification layer.

In the text encoder, we employ SciBERT as a text encoder to extract the text features that act as the input to the classification layer. The input of the text encoder can be represented as follows:

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

where $token_i$ denotes the i^{th} token of the input sentence $sent$ as tokenized by the corresponding tokenizer. m is the token number of $sent$. $[CLS]$ and $[SEP]$ correspond to the special symbol at the beginning and the end of



Query	how to imporve BLEU score
Task	Machine translation

Operation	Source sentence
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 .

Figure 6: Mechanism Knowledge Graph and Search Engine Construction Process.

the sentence, respectively. We can obtain the text vector representation h via SciBERT:

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

In the classification layer, we use h_{CLS}, which is the first component of h and corresponds to the $[CLS]$ token, as the input to the classification layer, which includes a dropout layer and a fully connected layer. Finally, we apply a softmax function to the label logits to obtain the probability distribution regarding whether the input sentence contains metrics-driven mechanism knowledge.

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

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

4.1.2. Mechanism Triple Extraction

As described in Section 3.1, the metrics-driven mechanism is a triple (*Operation, Effect, Direction*) triple, where the *Operation* and the *Effect* are entities, and the *Direction* is the relationship between the *Operation* and the *Effect*. Therefore, we formalize the mechanism triple extraction as an entity and relation extraction task. There are two types of approaches for entity and relation extraction tasks: pipeline-based approaches and end-to-end joint approaches. SpERT, proposed by Eberts and Ulges [27], is a state-of-the-art end-to-end joint entity and relation extraction method.

We finetune SpERT on our dataset to jointly extract entities and relations. As shown in Figure 6, SpERT first obtains the representation of span and classifies the entity category of span. Second, SpERT combines the entities in pairs to form the representation of relations between entities. Finally, entity pairs are classified as one of {*positive effect,negative effect, other*}.

The metrics-driven mechanism knowledge is extracted by employing the finetuned model from the abstracts of 26k ACL papers. In the extracted mechanism triples, some *Operation* entities are pronouns (e.g., "our model", "proposed method" and "new algorithm") instead of concrete entities. To alleviate the influence of this problem, we adopt the coreference resolution method proposed by AllenAI[46].

4.2. Construction of the NLP Mechanism KB

Task entities refer to research problem in NLP scientific papers. We further extend the mechanism triple $(\boldsymbol{Operation}\ ,\boldsymbol{Effect},\ \boldsymbol{Direction})$ into the $(\boldsymbol{Operation}, \boldsymbol{Effect}, \boldsymbol{Direction}, \boldsymbol{Task})$ n-ary mechanism relation. Therefore, the proposed NLP metrics-driven mechanism KG schema contains three types of entities: tasks, operations, and effects. According to the schema in Section 3.1, there are three relation types (*positive effect, negative effect*, and *other*) that describe the influence direction between the operation entity and the effect entity. In addition, we use evaluatedBy to describe the relation between the task entity and the effect entity.

The paper research task entity extraction problem is formalized as a multi-label classification task because of the uncontrollable research task extraction result based on the sequence-labeling approaches. In the Papers With Code(PWC)[3] , there is a taxonomy of tasks and subtasks [5]. In addition, there are many available papers with metadata that indicate the research areas, tasks or subtasks. Based on the BERT model and PWC dataset, a paper task classification model was finetuned, and it achieves an 87 F1 score.

For a paper without extraction $(Operation, Effect, Direction, Task)$ n-ary mechanism relation, we use the $(Method, Metric, Direction, Task)$ n-ary relation as a pseudo n-ary mechanism relation to express the knowledge that is similar to the metrics-driven mechanism knowledge in our work. Entities in pseudo n-ary mechanism relations are extracted by the SpERT[27] model trained on the SCERC[12] dataset. Note that the method entity and the metric entity refer to the Operation entity and effect entity, respectively. The direction between them is set as "unknown".

Finally, we build a knowledge base of metrics-driven mechanisms in the NLP domain (**NLPMKB**) that consists of 24k n-ary mechanism relations and 76k pseudo n-ary mechanism relations in the form of (*Method, Metric, Direction,Task*).

4.3. Construction of Mechanism Knowledge Search Engine

The **NLPMKB**enables applications to retrieve metrics-driven mechanisms in NLP. For example, a user can search all papers that contain a mechanism related to the question: *how to improve the diversity of the keyphrases extraction task.* To build the search engine of mechanism knowledge, we first use the multi-qa-MiniLM model[4], which maps sentence and query text to a 384 dimensional dense vector space. Then, we compute the cosine similarity score to find potentially relevant papers. Finally, we rerank the retrieved sentences using Cross-Encoder for MS Marco[5].

Table 4
Result of mechanisms sentence extraction

Type	Precision	Recall	F1 score
Non mechanism sent	92.5	93.0	92.7
Mechanism sent	77.8	76.7	77.2
Total	89.0	89.0	89.0

5. Evaluating The Extracted Mechanism Knowledge

In this section, we first evaluate the trained mechanism extraction model (Section 5.1). Then, we evaluate the quality of the extracted mechanism knowledge from the perspective of correctness and coverage (Section 5.2). Finally, we evaluate the utility of extracted metric-driven mechanism knowledge (Section 5.3) in terms of the mechanism knowledge search scenario.

5.1. Model Evaluations

Evaluation of Subtask 1

For subtask 1, as shown in Table 4, our mechanism sentence extraction model achieves an 89 F1 score on the testset, which has 454 non-mechanism sentences and 146 mechanism sentences.

Deep learning models are commonly referred to as black boxes. To understand the reasons underlying the decision making process and avoid avoid the detection of incorrect features in the data by the model, we adopt the Local Interpretable Model-agnostic Explanations (LIME), an explainable artificial intelligence (xAI) framework proposed by Ribeiro et al. [47], to interpret the mechanism sentence extraction model.

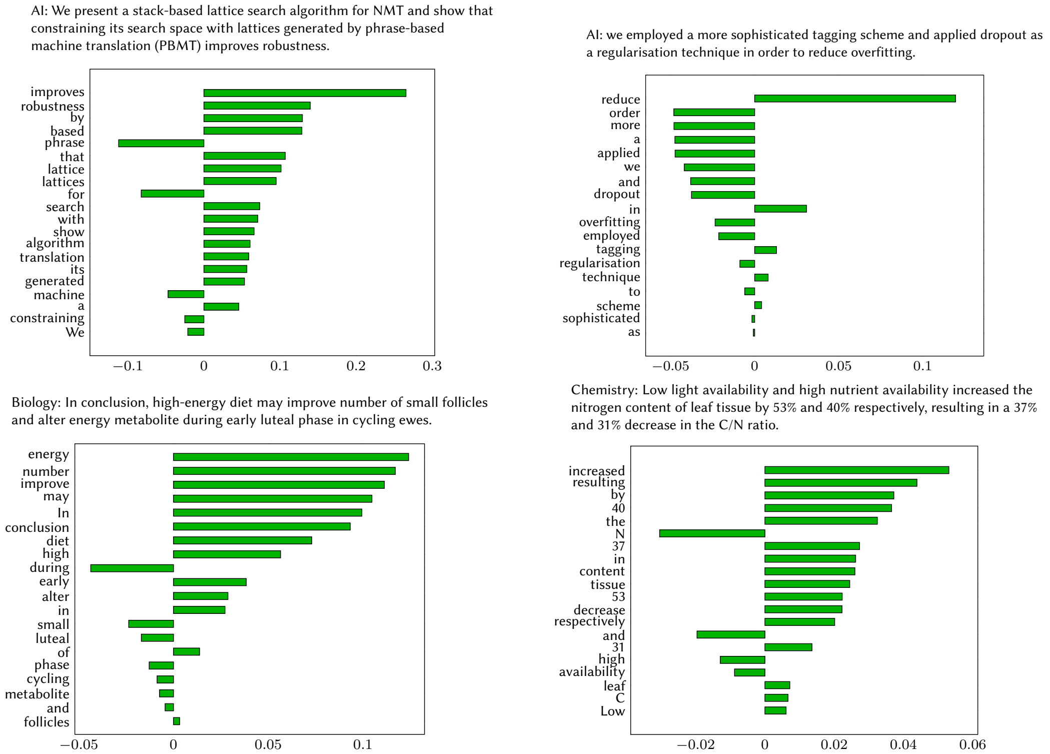
The LIME model is one of the most popular model-agnostic frameworks, and it primarily focuses on explaining individual predictions. As for the text classification task, LIME samples instances around an individual input text instance by adding a perturbation to the original text; one example of a perturbation involves randomly deleting words from the original text. Then, LIME classifies the generated samples using the trained model. Finally, the contribution of each word in the original text to the final model prediction result is obtained by the LIME framework.

We randomly select two sentences that contain metrics-driven mechanism knowledge, as shown in Fig-

[3] Papers With Code (PWC) is an open source repository about papers, datasets, and evaluation in the machine learning(ML) community and natural language processing(NLP) community created by researchers at Facebook AI Research. We downloaded the PWC dataset (licensed under CC BY-SA 4.0). We focus on the Papers with abstracts archive. https://paperswithcode.com

[4] https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

[5] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2



AI: We present a stack-based lattice search algorithm for NMT and show that constraining its search space with lattices generated by phrase-based machine translation (PBMT) improves robustness.

AI: we employed a more sophisticated tagging scheme and applied dropout as a regularisation technique in order to reduce overfitting.

Biology: In conclusion, high-energy diet may improve number of small follicles and alter energy metabolite during early luteal phase in cycling ewes.

Chemistry: 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.

Figure 7: Examples in artificial intelligence, biology and chemistry field for mechanism sentence extraction based on LIME framework

Table 5
Result of metrics-driven mechanism recognition

	Type	Precision	Recall	F1 score
Entities Extraction	Operation	72.0	66.4	69.1
	Effect	86.3	87.6	86.9
	Total	79.8	77.6	78.7
Relations Extraction	Pos_eff	59.7	71.7	65.1
	Neg_eff	50.0	54.2	52.0
	Other	60.0	26.1	36.4
	Total	58.1	61.6	59.8

ure 7, to make reasonable interpretations about our mechanism sentence extraction model. In Figure 7, the x-axis refers to the word contribution to the prediction result, where the positive and negative values correspond to the probability that the sentence contains mechanism knowledge or not, respectively. In Figure 7, the verb "improve" and "reduce" that represent the metric changes direction have the biggest influence on the prediction.

Benefiting from the BERT model’s strong ability in feature extraction and the domain generalizability of the metrics-driven mechanism representation scheme, although the dataset proposed in this paper is in the field of natural language processing, our mechanism sentence extraction model and mechanism triple extraction model still have good performance in other fields, such as biology and chemistry. the second row in Figure 7 demonstrates the decent generalization performance in other fields such as biology and chemistry.

Based on the LIME framework, it can be found that the BERT- based model primarily focuses on key verbs, such as increase, improve, reduce, and decrease, which indicate the metric entities change direction, to identify

Table 6
Human evaluation result for the enhanced mechanism knowledge search engine.

		Our search engine		Baseline	
No	Query	P@3	P@5	P@3	P@5
1	how to improve F1 on text classification?	100	60	33	60
2	how to improve model generalization?	100	80	100	60
3	how to decrease training time?	100	100	100	100
4	how to improve performance of Named Entity Recognition (NER)?	67	80	33	40
5	how to improve BLEU on machine translation?	100	100	100	100
	Avg.	**93**	**84**	73	72

Table 7
Human evaluation result of COLING 2020 papers

	Ground truth	
Predicted	Positive	Negative
Positive	27	4
Negative	7	62

the sentences that contain the metrics-driven mechanism knowledge.

Evaluation of Subtask 2

Identical to the scientific information extraction in terms of entity granularity, our mechanism triple extract model achieves a 78.67 F1 score on both Operation and Effect entity recognition. For relation extraction, it achieves a 59.80 F1 score. Using the same method, our mechanism entity and relation extraction model outperforms SCIERC, which achieves a 70.33 and 50.84 F1 score corresponding to entity recognition and relation extraction, respectively.

5.2. Quality Evaluation of Extracted Mechanism Knowledge.

To evaluate the quality of the metrics-driven mechanism knowledge extracted from the paper abstracts, we randomly selected 100 papers in 2020 COLING and checked the extracted (*Operation, Effect, Direction*) triples with a relaxed-match evaluation [48], i.e., an entity is regarded as positive if its type is correct and there is an overlap with the ground truth entity boundary.

There are 34 papers that contain metrics-driven mechanism knowledge in their abstracts. The confusion matrix is shown in Table 7. We achieve 87.0 precision and 79.4 recall. According to the analysis of seven papers, the mechanism knowledge could not be extracted, which we find was caused by the error cascade in the mechanism's knowledge extraction model. For instance, "Our word segmentation system outperforms the previous state-of-the-art system in both speed and accuracy on both in-domain and out-domain datasets." actually contains the mechanism knowledge, but the mechanism sentence extraction model fails to recognize it.

5.3. Utility Evaluation of Extracted Mechanism Knowledge

Using the PWC hierarchical task taxonomy, our NLP Mechanism KB supports the automatic semantic extension of tasks such as extending *Text Generation* to *Paraphrase Generation*, *News Generation* and *Paper generation.* Therefore, for a query about *Text Generation*, our NLP Mechanism KB can return mechanism knowledge for *Paraphrase Generation*, *News Generation* and *Paper generation.*

To illustrate the utility of our mechanism knowledge search engine, we map the text (e.g., the abstract sentences in a paper, research tasks, and input query) to a shared vector space $\mathbb{R}^d$, where d is the vector dimension. In the similarity calculation step, a abstract sentence and research task are concatenated together to obtain the semantic vector. Then, the cosine similarity score between the two semantic vectors is calculated. In the evaluation, our constructed mechanism knowledge search engine is compared with the baseline without mechanism knowledge enhancement, which uses all of the sentences in an abstract as potential candidates instead of the extracted mechanism sentence. For the sake of fairness, the baseline uses the same similarity calculation method and backbone ranking model as our search engine.

As shown in Table 6, the enhanced mechanism knowledge search engine achieves a significantly better performance against the baseline method. Specifically, in terms of P@3 and P@5, the enhanced mechanism knowledge search engine could achieve 20- and 12-points improvements compared with the baseline method, respectively.

6. Conclusion

In this paper, we introduce a coarse-grained representation scheme to express metrics-driven mechanisms in the field of artificial intelligence. Our scheme achieved a balance between domain adaptability and universality. Moreover, we construct a dataset based on the abstracts of papers in the NLP field for mechanism sentence extraction and mechanism triple extraction. Based on the annotated dataset, a BERT-based metric-driven mechanism knowledge extraction model is trained and a knowledge base of metrics-driven mechanism in the NLP field is constructed. The human evaluation shows that our metrics-driven mechanism knowledge base has high quality, and the extracted mechanism knowledge achieves 87.0 precision and 79.4 recall. Additionally, we find that the mechanism search performance is improved by using the extracted metrics-driven mechanism knowledge.

Benefitting from the pre-trained model's learning ability and the domain generalizability of the metrics-driven mechanism representation scheme proposed in this paper, the trained model also has the ability to extract metrics-driven mechanism knowledge in the fields of biology and chemistry. In the future, we will extract metrics-driven mechanism knowledge distributed in multiple sentences and explore the few-shot learning method to build a mechanism extraction model for general fields.

References

[1] P. Machamer, L. Darden, C. F. Craver, Thinking about Mechanisms, Philosophy of Science 67 (2000) 1–25.

[2] S. Glennan, Mechanisms and the nature of causation, Erkenntnis 44 (1996). doi:10.1007/BF00172853.

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

[4] F. Martínez-Plumed, P. Barredo, S. hÉigeartaigh, J. Hernández-Orallo, Research community dynamics behind popular AI benchmarks, Nature Machine Intelligence 3 (2021) 581–589. doi:10.1038/s42256-021-00339-6, number: 7 Publisher: Nature Publishing Group.

[5] B. Koch, E. Denton, A. Hanna, J. G. Foster, Reduced, reused and recycled: The life of a dataset in machine learning research, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021.

[6] D. Schlangen, 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), Association for Computational Linguistics, Online, 2021, pp. 670–674. doi:10.18653/v1/2021.acl-short.85.

[7] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, M. Gardner, A dataset of information-seeking questions and answers anchored in research papers, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4599–4610. doi:10.18653/v1/2021.naacl-main.365.

[8] D. Zhang, S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. C. Niebles, M. Sellitto, et al., The ai index 2021 annual report, ArXiv preprint abs/2103.06312 (2021). URL: https://arxiv.org/abs/2103.06312.

[9] M. Färber, A. Albers, F. Schüber, Identifying used methods and datasets in scientific publications., in: SDU@ AAAI, 2021.

[10] Y. Hou, C. Jochim, M. Gleize, F. Bonin, D. Ganguly, 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, Association for Computational Linguistics, Online, 2021, pp. 707–714. doi:10.18653/v1/2021.eacl-main.59.

[11] S. Jain, M. van Zuylen, H. Hajishirzi, I. Beltagy, SciREX: A challenge dataset for document-level information extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7506–7516. doi:10.18653/v1/2020.acl-main.670.

[12] Y. Luan, L. He, M. Ostendorf, H. Hajishirzi, Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3219–3232. doi:10.18653/v1/D18-1360.

[13] Y. Wang, C. Zhang, Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing, Journal of Informetrics 14 (2020) 101091. doi:https://doi.org/10.1016/j.joi.2020.101091.

[14] R. Islamaj, R. Leaman, S. Kim, D. Kwon, C.-H. Wei, D. C. Comeau, Y. Peng, D. Cissel, C. Coss, C. Fisher,

et al., Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature, Scientific Data 8 (2021) 1–12.

[15] M. E. Savery, W. J. Rogers, M. Pillai, J. G. Mork, D. Demner-Fushman, Chemical entity recognition for medline indexing, AMIA Summits on Translational Science Proceedings 2020 (2020) 561.

[16] N. Greenberg, T. Bansal, P. Verga, A. McCallum, Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2824–2829. doi:10.18653/v1/D18-1306.

[17] Z. Nasar, S. W. Jaffry, M. K. Malik, Information extraction from scientific articles: a survey, Scientometrics 117 (2018) 1931–1990. doi:10.1007/s11192-018-2921-5, 00000.

[18] V. Z. Chen, F. Montano-Campos, W. Zadrozny, Causal knowledge extraction from scholarly papers in social sciences, ArXiv preprint abs/2006.08904 (2020). URL: https://arxiv.org/abs/2006.08904.

[19] T. Hope, A. Amini, D. Wadden, M. van Zuylen, S. Parasa, E. Horvitz, D. Weld, R. Schwartz, H. Hajishirzi, Extracting a knowledge base of mechanisms from COVID-19 papers, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4489–4503. doi:10.18653/v1/2021.naacl-main.355.

[20] W. Bechtel, The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar, Erkenntnis 73 (2010) 313–328. doi:10.1007/s10670-010-9234-2.

[21] J. Röhl, Mechanisms in biomedical ontology, Journal of Biomedical Semantics 3 (2012) S9. doi:10.1186/2041-1480-3-S2-S9, 00016.

[22] F. Yang, L. G. Moss, G. N. Phillips, The molecular structure of green fluorescent protein, Nature biotechnology 14 (1996) 1246–1251.

[23] Y. S. Chan, D. Roth, Exploiting syntactico-semantic structures for relation extraction, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 551–560.

[24] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, 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), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 2124–2133. doi:10.18653/v1/P16-1200.

[25] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 50–61. doi:10.18653/v1/2021.naacl-main.5.

[26] D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, relation, and event extraction with contextualized span representations, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5784–5789. doi:10.18653/v1/D19-1585.

[27] M. Eberts, A. Ulges, Span-based Joint Entity and Relation Extraction with Transformer Pre-training, ArXiv preprint abs/1909.07755 (2019). URL: https://arxiv.org/abs/1909.07755.

[28] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, X. Qiu, 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), Association for Computational Linguistics, Online, 2021, pp. 5808–5822. URL: https://aclanthology.org/2021.acl-long.451. doi:10.18653/v1/2021.acl-long.451.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, 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, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[30] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, J. Li, Entity-relation extraction as multi-turn question answering, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1340–1350. doi:10.18653/v1/P19-1129.

[31] S. N. Kim, O. Medelyan, M.-Y. Kan, T. Baldwin, SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 21–26.

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

[33] W. Lu, P. Li, G. Zhang, Q. Cheng, 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 (2020) 1320. doi:10.3772/j.issn.1000-0135.2020.12.008.

[34] J. D'Souza, A. Hoppe, A. Brack, M. Y. Jaradeh, S. Auer, R. Ewerth, 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, European Language Resources Association, Marseille, France, 2020, pp. 2192–2203.

[35] P. Li, Q. Liu, Q. Cheng, W. Lu, Data set entity recognition based on distant supervision, The Electronic Library (2021).

[36] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 546–555. doi:10.18653/v1/S17-2091.

[37] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, T. Charnois, SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 679–688. doi:10.18653/v1/S18-1111.

[38] I. Mondal, Y. Hou, C. Jochim, End-to-end nlp knowledge graph construction, ArXiv preprint abs/2106.01167 (2021). URL: https://arxiv.org/abs/2106.01167.

[39] T. Hope, J. Chan, A. Kittur, D. Shahaf, Accelerating innovation through analogy mining, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, ACM, 2017, pp. 235–243. URL: https://doi.org/10.1145/3097983.3098038. doi:10.1145/3097983.3098038.

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

[41] S. Teufel, A. Siddharthan, C. Batchelor, Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2009, pp. 1493–1502.

[42] P. Accuosto, H. Saggion, Mining arguments in scientific abstracts with discourse-level embeddings, Data & Knowledge Engineering 129 (2020) 101840. doi:10.1016/j.datak.2020.101840.

[43] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv preprint abs/1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[45] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[46] K. Lee, L. He, L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 687–692. doi:10.18653/v1/N18-2108.

[47] M. T. Ribeiro, S. Singh, C. Guestrin, "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, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[48] R. Grishman, B. Sundheim, Message Understanding Conference- 6: A brief history, in: COLING 1996 Volume 1: The 16th International Conference on

Computational Linguistics, 1996.