

Exploring developments of the AI field from the perspective of methods, datasets, and metrics

Rujing Yao^{a,b}, Yingchun Ye^b, Ji Zhang^c, Shuxiao Li^d and Ou Wu^{b,*}

^aDepartment of Information Resources Management, Business School, Nankai University, Tianjin, 300071, China

^bCenter for Applied Mathematics, Tianjin University, Tianjin, 300072, China

^cInstitute of AI, Zhejiang Lab, Hangzhou, 311121, China

^dInstitute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Keywords:

AI literature

Named entity recognition

Self-paced learning

Entity-level analysis

ABSTRACT

The knowledge contained in academic literature is interesting to mine. Inspired by the idea of molecular markers tracing in the field of biochemistry, three named entities, namely, methods, datasets, and metrics, are extracted and used as artificial intelligence (AI) markers for AI literature. These entities can be used to trace the research process described in the bodies of papers, which opens up new perspectives for seeking and mining more valuable academic information. Firstly, the named entity recognition model is used to extract AI markers from large-scale AI literature. A multi-stage self-paced learning strategy (MSPL) is proposed to address the negative influence of hard and noisy samples on the model training. Secondly, original papers are traced for AI markers. Statistical and propagation analyses are performed based on the tracing results. Finally, the co-occurrences of AI markers are used to achieve clustering. The evolution within method clusters is explored. The above-mentioned mining based on AI markers yields many significant findings. For example, the propagation rate of the datasets gradually increases over time. The methods proposed by China in recent years have an increasing influence on other countries.

1. Introduction

The literature in the subject of artificial intelligence (AI) has significantly increased with the field's rapid progress, and research in the field is having an increasing influence on a variety of fields (Ma et al., 2022b). Sorting out the overall context of the AI field and exploring the characteristics of its development are crucial for individuals engaged in AI research, beginners in AI research, and researchers in other disciplines affected by AI. The exploration of AI academic literature can help these researchers quickly and accurately seek research information and understand research trends. With the evolution of various methods, datasets, and metrics, their extraction and analysis in AI literature can help researchers to quickly understand the application and development of these entities, and selecting appropriate entities for their own research.

In the biological field, molecular markers are frequently used to track the changes in the substances and cells during the reaction to obtain reaction characteristics and regularities (Zhao et al., 2018a, 2019). For example, Ruben et al. (1941) used the isotope of oxygen ^{18}O to mark H_2O and CO_2 and track the source of O_2 in photosynthesis. Inspired by this, we observe that methods, datasets, and metrics can play the same role as molecular markers in AI literature mining. As shown in Fig. 1, when these entities are proposed or cited by different literature, the traces in the specific research process are formed. Accordingly, the methods, datasets, and metrics, which are in the same granularity, in the AI literature are used as AI markers. These factors can be leveraged to trace the information reflecting the research process in the paper bodies. Given that abstracts mainly contain conclusive information and lack information reflecting the research process, and the bodies of the papers provide the specific process of research, we extract AI markers from the paper bodies.

In recent years, an increasing number of research has focused on the analysis of the key entities in the AI literature, such as methods, datasets, and metrics (Wang and Zhang, 2020; Li et al., 2021b; Zhang et al., 2021b). To our knowledge, there is no study that simultaneously extracts and utilizes these three entities for the analysis of the AI field. In this

*Corresponding author

 rjyao@mail.nankai.edu.cn (R. Yao); yingchunye@tju.edu.cn (Y. Ye); zhangji77@gmail.com (J. Zhang); shuxiao.li@ia.ac.cn (S. Li); wuou@tju.edu.cn (O. Wu)

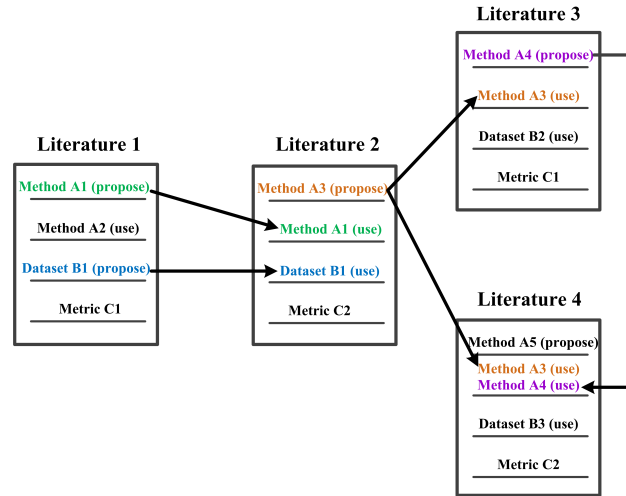


Figure 1: Traces of the AI markers proposed or cited by different literature.

study, large-scale AI papers are collected. The three entities are extracted and linked to analyze the AI field. In addition, a multi-stage self-paced learning strategy (MSPL) is proposed to address the negative influence of hard and noisy samples on extraction model training.

In summary, our contributions are as follows:

- 1) Large-scale AI papers are collected. Method, dataset, and metric entities are extracted from the bodies of the collected papers.
- 2) We introduce MSPL to address the negative influence of hard and noisy samples on extraction model training. This strategy assigns different weights to distinct training samples according to their learning difficulties in various training stages. To the best of our knowledge, this work is the first to adapt the idea of SPL in named entity recognition (NER). The experiments show that the proposed MSPL strategy improves the performance of the entity extraction model.
- 3) Method, dataset, and metric entities are used as AI markers to explore the development of the AI field, and these entities are linked for analysis. Numerous significant findings are achieved. The details are as follows.

- The annual development of the AI field is obtained on the basis of the extracted AI markers (methods and datasets). For example, MNIST (LeCun et al., 1998) ranked first in terms of usage from 2005 to 2014. ImageNet (Deng et al., 2009) ranked first after 2014.
- Based on the original paper tracing for AI markers, the United States, China, and the United Kingdom are relatively active countries in the field of AI. The propagation rate of the datasets gradually increases over time. Based on the propagation analysis of the methods among countries, the methods proposed by China have an increasing influence on other countries.
- Method roadmaps are constructed on the basis of the method clusters and associated datasets, which can show the evolution in method clusters.

2. Related work

This study involves several aspects, including extraction and bibliometrics on knowledge entities, structure function identification, NER, and term function recognition.

2.1. Extraction and bibliometrics on knowledge entities

The existing research on knowledge entities includes a number of aspects, such as methods, datasets, metrics, and software (Ding et al., 2013b; Zhang et al., 2021a). Zhang et al. (2021a) used a dictionary-based approach to identify methods in academic papers. Wang and Zhang (2020) manually annotated the algorithm entities in the full text of ACL

conference papers from 1979 to 2015. A dictionary-based approach was used by Ding et al. (2019) to identify algorithms from the full text of ACL papers. Wang and Zhang (2018) utilized a dictionary to extract algorithms in papers published in ACL. CRF and BiLSTM+CRF were used to extract model entities from academic papers (Lei and Wang, 2019). To recognize dataset entities in academic papers, a method based on distant supervised learning was proposed by Li et al. (2021b), which is a pioneering study to apply distant learning to dataset recognition. Heddes et al. (2021) used SciBERT to extract datasets from scientific articles. A metrics-driven mechanism knowledge representation schema was proposed by Ma et al. (2022b), and the SpERT method was used to jointly extract entities and relations from the abstracts of AI papers. Zhang et al. (2021b) manually annotated the metric, tool, resource and method entities from full text of papers published in China National Conference on Computational Linguistics. To extract software entities from full-text papers, an improved bootstrapping method was proposed (Pan et al., 2015).

Taking the NLP field as an example, Zhang et al. (2021a) compared the mention of algorithms in full-text papers from an English conference and a Chinese conference. The comparative analysis includes frequency, location, and time of the algorithms. Wang and Zhang (2020) proposed an equation to measure the influence of an algorithm. The influence of different algorithms in ACL papers, top 10 algorithms in different ages, and the evolution of influence of various algorithms were analyzed. The citation frequency and citation time evolution of algorithms were revealed by Ding et al. (2019). The comparison of the top 10 data mining algorithms in terms of the number of papers, frequency of algorithms mentioned, and location of algorithms were conducted. In addition, the task of each article is classified and the most relevant task for each algorithm is obtained (Wang and Zhang, 2018). Zhao et al. (2018b) explored mentions and citations of datasets in a multidisciplinary full-text corpus. The overview of data mentions and citations was demonstrated from multiple perspectives, such as sections where datasets are mentioned, trackability of datasets, and types of data archives. The disciplinary characteristics were also analyzed from many aspects, such as the percentage of article authors collected data themselves in each discipline. Zhang et al. (2021b) constructed an association network among metric, tool, resource, and method entities to explore the relevance. The top 50 most frequently used software and most highly cited software were reported by Pan et al. (2015). Furthermore, they explored the relationship between the number of mentions and the number of citations. Zhang et al. (2019) extracted software from the full text of academic articles in PLOS ONE. The cluster analysis was conducted to explore the connections among scientific software, and the top five clusters with the largest number of software were shown.

2.2. Structure function identification

The identification of structure function in academic literature has attracted the attention of many scholars. Rule-based methods and machine learning-based methods are the main methods. Kim et al. (2000) identified the title, author, affiliation, and abstract in academic literature based on 120 rules. Ding et al. (2013a) used section orders and keywords to identify the structure function of academic papers, and studied the citation distribution of different sections in a paper. A rule-based system was proposed to identify the structure function of academic articles (Constantin et al., 2013). With the development of machine learning, an increasing number of researchers use machine learning methods to identify structure functions. Tuarob et al. (2015) used machine learning algorithms to identify the section boundaries of academic literature. SVM, Text-CNN, and BERT were used to identify the structure function of academic text (Ji et al., 2019). Ma et al. (2021) employed a variety of classification models to identify the structure function of academic literature, and effective characteristics were screened out through experiments. Lu et al. (2018) refined the identification of structure function into section header, section content, and paragraph. A novel clustering algorithm was proposed to generate the structure function of a specific domain. Ma et al. (2022a) used traditional machine learning methods and deep learning methods to train the structure function identification model.

2.3. NER

NER is a crucial task in natural language processing (NLP) and has a wide range of applications, such as question answering (Mollá et al., 2006), machine translation (Siekmeier et al., 2021), and information extraction (Derczynski et al., 2015). The methods for NER mainly include rule-based methods and machine learning-based methods.

Rule-based methods manually construct rule templates or dictionaries to match named entities. Hand crafted lexical resources were exploited to construct a rule-based Greek NER system (Farmakiotou et al., 2000). Riaz (2010) exploited a rule-based method for NER in Urdu, which outperforms statistical learning models. With the vigorous development of machine learning, machine learning-based methods have also been used to extract entities in the literature. A feature generation method using features of complex SRs and simple SRs is proposed, and it can improve the performance of NER (Cho et al., 2013). Sentence-level and document-level representations were used by Luo et al. (2020) to

augment the performance of NER. Recently, the main research directions of NER are few-shot learning, nested NER, and discontinuous NER. A self-describing mechanism is proposed to solve the challenges (limited information and knowledge mismatch) in few-shot NER. SDNet is also proposed, and a universal knowledge can be obtained by SDNet (Chen et al., 2022). Retrieval-based span-level graphs were used to link spans and entities to obtain a better span representation in nested NER (Wan et al., 2022). A span-based model was proposed by Li et al. (2021a) to recognize discontinuous and overlapped entities.

2.4. Term function recognition

Term function recognition is an information extraction task. Kondo et al. (2009) identified heads, methods, and goals from research papers' titles. An unsupervised bootstrapping algorithm is proposed to recognize technologies and applications from scientific literature (Tsai et al., 2013). Li et al. (2017) designed a novel literature analysis system named CS-LAS, which can semantically analyze scientific literature from the perspective of term function recognition. Nanba et al. (2010) considered the extraction of technologies and effects as a sequence-labeling problem, and extracted them from research papers and patents. Lu et al. (2019) manually annotated the term function of author-selected keywords, and analyzed the patterns of author-selected keywords. Cheng et al. (2021) transformed the information extraction task into a specific form of title generation. They proposed a novel method combining a deep learning and title generation strategy to recognize term function. Lu et al. (2020) designed an effective supervised neural network method to achieve recognition of the research questions and methods in academic texts. A novel system, AKMiner, was proposed to extract method and task concepts from academic literatures (Huang and Wan, 2013). New supervised classifiers were proposed to extract problems and solutions from scientific texts (Heffernan and Teufel, 2018).

3. Data

A large amount of AI literature is necessary for our study. Firstly, this section introduces the literature data we collected. In addition, two machine learning models are used during the research. Therefore, the section also introduces the training data of these two models.

3.1. Collected literature data

A total of 122,446 papers published from 2005 to 2019 were collected by using the list of AI journals and conferences in China Computer Federation (CCF)¹ ranks (Tier-A, Tier-B, and Tier-C). The number of papers collected at each publication venue and the websites for collection are shown Table A1 in Appendix. GROBID² is utilized to convert PDF format papers into XML format. The data is obtained by extracting certain pieces of information, including titles, countries, publication venues, years, bodies, and references, from the papers in XML format. To facilitate reading, the collected data is called CCF corpus.

3.2. Training data for the chapter classification

The main body of an AI paper generally includes four chapters: introduction, methodology, experiment, and conclusion. The roles of AI markers in different parts vary. This study introduces a chapter classification strategy to divide the body of AI literature into the above-mentioned four parts.

A total of 2000 papers in the CCF corpus are randomly selected to train the chapter classifier. The data labeling process is as follows. Firstly, ten graduate students engaged in AI research are recruited to label the data. The ten annotators are divided into five groups. In each group, two annotators are asked to independently label the same data. Secondly, after labeling, we measure the interrater reliability (IRR) between the two annotators in the same way as Wang and Zhang (2020), and the IRR is 0.90, which shows that our annotations are sufficiently reliable. Thirdly, the two annotators discuss the differences in the labeling data to determine the final labeling results. Finally, we recruit another five graduate students engaged in AI research to conduct a thorough check of the labeled data. When at least four annotators think that the label of a sample is wrong, the label is corrected.

This data corpus is called TCCdata and is used to construct a BiLSTM classifier for chapter classification. The numbers of chapters and the associated numbers of paragraphs for each chapter in TCCdata are shown in Table 1.

¹CCF compiled a list of AI journals and conferences with different ranks. See <https://www.ccf.org.cn/en/Bulletin/2019-05-13/663884.shtml> for details.

²<https://grobid.readthedocs.io/en/latest/>

Table 1

Numbers of chapters and paragraphs in the TCCdata.

	Chapters	Paragraphs
Introduction	2918	24,385
Methodology	1004	14,344
Experiment	1961	18,289
Conclusion	2391	6092

Table 2

Numbers of AI markers in the TMEdata.

	AI markers		
	Method	Dataset	Metric
Training	4737	2526	1046
Validation	815	528	205
Testing	567	298	159

3.3. Training data for the AI marker extraction

A total of 1000 papers from the CCF corpus are randomly selected to learn an AI marker extraction model. The methodology and the experiment chapters of the 1000 papers are divided into sentences according to punctuation. The BIO labelling strategy (Ratinov and Roth, 2009) is adopted for the three AI markers, namely, methods, datasets, and metrics. The methods, datasets, and metrics compiled by JiqiZhixin³ are used to pre-annotate our compiled data, which can accelerate the human labeling. The human labeling process of this data is the same as that of the TCCdata. The IRR is 0.81, indicating that our annotations are sufficiently reliable. Finally, 10,410 labelled sentences are obtained and are called TMEdata.

During the training of the extraction model, the TMEdata is divided into training, validation and testing sets according to the ratio of 7.5:1.5:1. The details are shown in Table 2.

4. Methodology

Fig. 2 depicts the framework of our work. After the bodies of the paper are obtained in xml format, we firstly classify the chapters, and then the sentences of the methodology chapter and experiment chapter are obtained. Subsequently, the sentences are sent to the AI marker extraction model to identify AI markers (i.e., methods, datasets, and metrics). Thereafter, the original paper tracing and clustering for AI markers are conducted. Finally, the statistical, propagation, and cluster analyses are used to reveal the development of the AI field. The source code is available at <https://github.com/researchondata/Mining-AI-entities>.

4.1. Chapter classification

In the body of an AI paper, the AI markers located in the methodology and the experiment chapters play a substantial role in the paper. Accordingly, only the AI markers of the methodology and the experiment chapters are extracted. Simple rule strategies are difficult to use to accurately classify the chapters of the AI literature due to the diversity of the structure of the AI literature. To train the structure function identification model, traditional machine learning methods and deep learning methods were utilized by Ma et al. (2022a). The experimental results showed that the BiLSTM hierarchical network is the most robust model to extract the features of chapter content among the competing methods. Furthermore, in numerous recent studies, BiLSTM is a frequently adopted model because of its good feature representation capability (Zhang et al., 2022; Adhikari et al., 2019). Therefore, to improve accuracy and efficiency, the chapter classification strategy that combines the BiLSTM algorithm and rules is adopted.

³<https://www.jiqizhixin.com/sota>

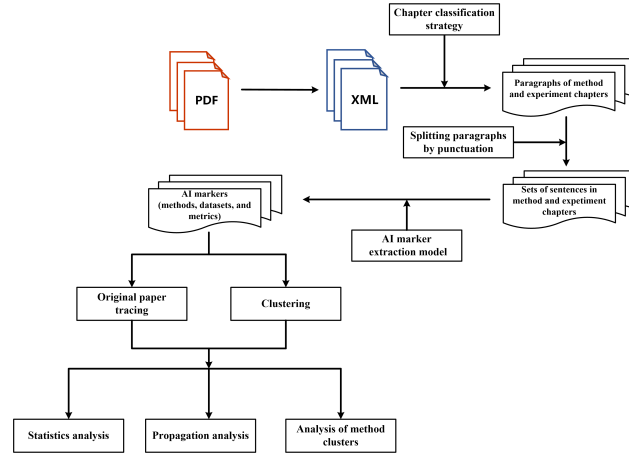


Figure 2: Framework of our work.

4.1.1. Proposed classification strategy

The overall of our strategy is shown in Fig. 3. Rules (e.g. keywords and orders) are firstly used to label the chapters. In well-matched chapters, the chapter labels are outputted. In unmatched chapters, the paragraphs under the chapters are inputted into the paragraph-level BiLSTM classifier trained based on the TCCdata for prediction. Next, the paragraph-level predicted labels in the same chapter are voted, and the labels with the most votes are used as the final label. Finally, the rule-based and BiLSTM-based results are combined to obtain the final chapter labels of the whole body.

The conventional one-layer BiLSTM architecture is adopted. The dimension of the word vector is 200, the hidden dimension is 256, and the batch size is 64. Cross entropy is used as the loss function. TCCdata is used as the training data.

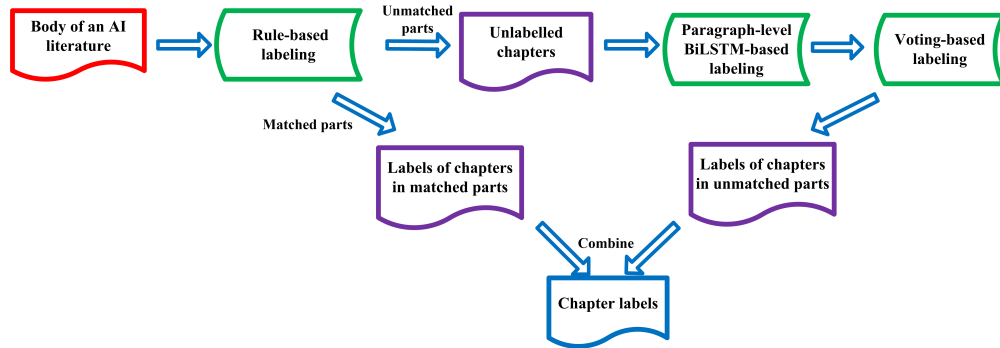


Figure 3: Overall process of chapter classification.

4.1.2. Evaluation results

The TCCData is divided into training, validation, and testing sets according to the ratio of 8:1:1. Three methods, namely, rule-based, BiLSTM-based, and combining-based, are evaluated. The accuracy is 0.7980 by only using rule matching. The accuracy is 0.7962 by only using paragraph-level BiLSTM trained on TCCData. The accuracy reached 0.9351 by using the combination.

4.2. AI marker extraction and normalization

The extraction and normalization of AI markers are challenging. Given that a large number of AI literature emerge every year, the number of new AI markers continues to increase, and the forms vary. No prescribed standard is imposed for the naming of AI markers. Some common words may also be used as datasets, such as ‘DROP’ in Dua et al. (2019).

4.2.1. AI marker extraction model

(1) CNN-BiLSTM-SA-CRF

AI marker extraction is a typical NER task (Shang and Ran, 2022). CNN-BiLSTM-CRF is a classic NER model that has been widely used in numerous studies with good performance (Shang and Ran, 2022; Dai et al., 2019). Self-attention (Vaswani et al., 2017) can capture the semantic features among words well. Therefore, the network structure of the AI marker extraction model is based on the CNN-BiLSTM-CRF model and self-attention is used in our extraction model. SciBERT (Beltagy et al., 2019) is a pretrained language model for scientific text and achieves good performance on many tasks, including scientific entity extraction. Therefore, SciBERT is used to obtain word-level embedding. The model structure is shown in Fig. 4.

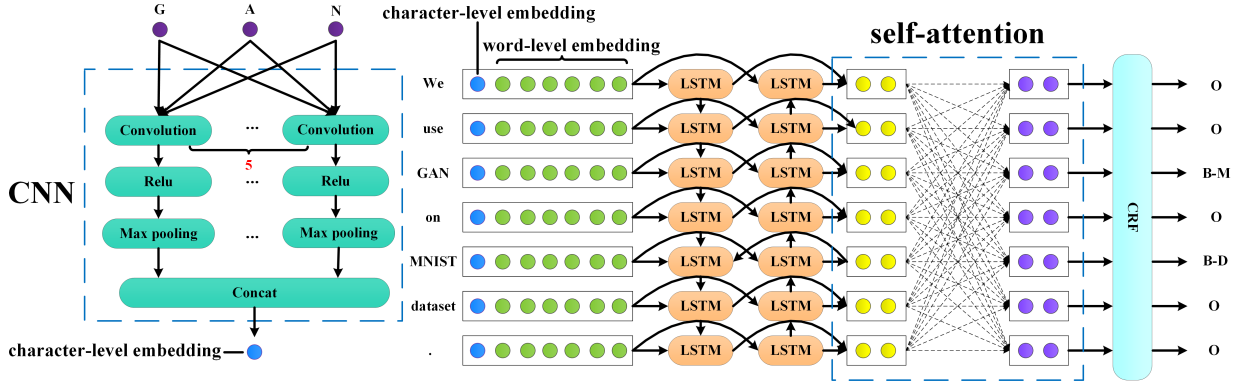


Figure 4: Structure of the AI marker extraction model.

Firstly, the character-level embedding for each word, w_i^{ch} , is obtained through a CNN network. Secondly, the word-level embedding for each word, w_i^{wd} , is obtained through SciBERT (Beltagy et al., 2019). Then, these two embeddings are concatenated as follows:

$$w_i^{mix} = [w_i^{ch}, w_i^{wd}]. \quad (1)$$

Subsequently, w_i^{mix} is fed to BiLSTM to capture contextual information.

$$\vec{h}_i = LSTM(\vec{h}_{i-1}, w_i^{mix}), \quad \tilde{h}_i = LSTM(\tilde{h}_{i+1}, w_i^{mix}), \quad (2)$$

where $\vec{h}_i$ and $\tilde{h}_i$ are the hidden vectors for the i th word obtained by forward LSTM and backward LSTM, respectively.

Then, $\vec{h}_i$ and $\tilde{h}_i$ are concatenated as follows:

$$h_i = [\vec{h}_i, \tilde{h}_i]. \quad (3)$$

Let $H = \{h_i\}_{i=1}^L$, where L is the (fixed) sentence length. Next, SA (Vaswani et al., 2017) is used to calculate the association among words. The final representation for each word is calculated as follows:

$$Q = HW^Q, K = HW^K, V = HW^V, \tilde{H} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (4)$$

where W^Q , W^K , and W^V are learnable parameters, and d is the length of h_i .

Finally, $\tilde{H}$ is fed to the CRF (Lafferty et al., 2001) to obtain the label sequence of the sentence.

(2) Multi-stage self-paced learning (MSPL)

Similar to other NER tasks (Zhu and Li, 2022; Wang et al., 2021), the recognition of AI named entities still meets the challenges, such as limited context and vague boundary. Moreover, noisy labels are nearly inevitable in NER

benchmark corpus (Wang et al., 2019; Li et al., 2020; Jie et al., 2019). Consequently, some training samples are more difficult to learn than others. To alleviate the negative influence of some quite difficult training samples, an easy-first learning strategy, namely, self-paced learning (SPL) (Kumar et al., 2010), is utilized. SPL simulates the mechanism of human learning, in which easy knowledge is learned first, followed by hard knowledge. In SPL, the samples are assigned different weights according to their difficulties. In the early training stage, the weights of easy samples (i.e., those with small losses below a threshold) are set to one, whereas the weights of hard samples (i.e., those with large losses) are set to zero. An increasing number of hard samples are involved in the training with the increase in epoch. The optimization with SPL is shown in Eq. (5).

$$\begin{aligned} \min \quad & \sum_i v_i l_i \\ \text{s.t.} \quad & v_i = \begin{cases} 1 & \text{if } l_i < \frac{1}{K} \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \quad (5)$$

where v_i is the weight of the i th sample, l_i is the loss of the i th sample, K is a hyper-parameter, and the value of K is iteratively reduced in the experiments. When the loss of a sample is less than $1/K$, v_i is equal to 1, which means that this sample is involved in training.

SPL adopts an easy-first weighting strategy. Nevertheless, another opposite strategy, namely, hard-first weighting, such as Focal Loss (Lin et al., 2017), is also widely leveraged in many deep learning tasks. Numerous experiments and applications verified the effectiveness of the hard-first weighting strategy. Motivated by these two apparently contradict weighting strategies, a multi-stage SPL strategy (MSPL) is proposed, which consists of the following three training stages:

- Stage 1. This stage adopts the easy-first strategy to alleviate the negative influence of the hard samples. Easy samples (i.e., samples with low training losses) are still assigned with high weights, and hard samples (i.e., samples with high training losses) are still with low weights. Accordingly, when the current epoch t is smaller than a threshold t_1 , the weights v_i are defined as follows:

$$v_i = \frac{1}{1 + \exp[\alpha(l_i - \tau)]}, \quad (6)$$

where τ is a hyperparameter.

- Stage 2. Equality stage. This training stage adopts the conventional strategy that all training samples have equal weights (the values are one). Accordingly, when the current epoch t locates in $[t_1, t_2]$, the weights v_i are defined as follows:

$$v_i \equiv 1. \quad (7)$$

- Stage 3. This stage adopts the hard-first stage to explore the potentiality of hard samples. Hard samples (i.e., samples with relatively high training losses) are assigned with high weights, and easy samples (i.e., samples with relatively low training losses) are with low weights.

$$v_i = \frac{1}{1 + \exp[-\alpha(l_i - \tau)]}. \quad (8)$$

In our experiments, the average loss of the current and the last two epochs is utilized to replace l_i in Eqs. (6) and (8) to increase the robustness.

To illuminate the difference between SPL and our proposed MSPL, we plot the weighting curves of SPL and MSPL, as shown in Figs. 5 and 6, respectively. MSPL can implement three various weighting strategies during different training stages.

4.2.2. Experiments

(1) Competing methods

The following methods are compared in the experiments to select a high-performance NER model for the extraction of methods, datasets, and metrics from large-scale AI literature.

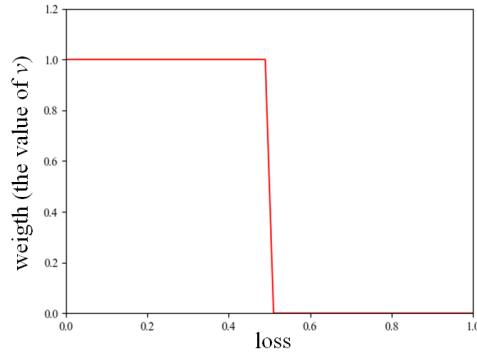


Figure 5: Curve of SPL when the threshold is 0.5.

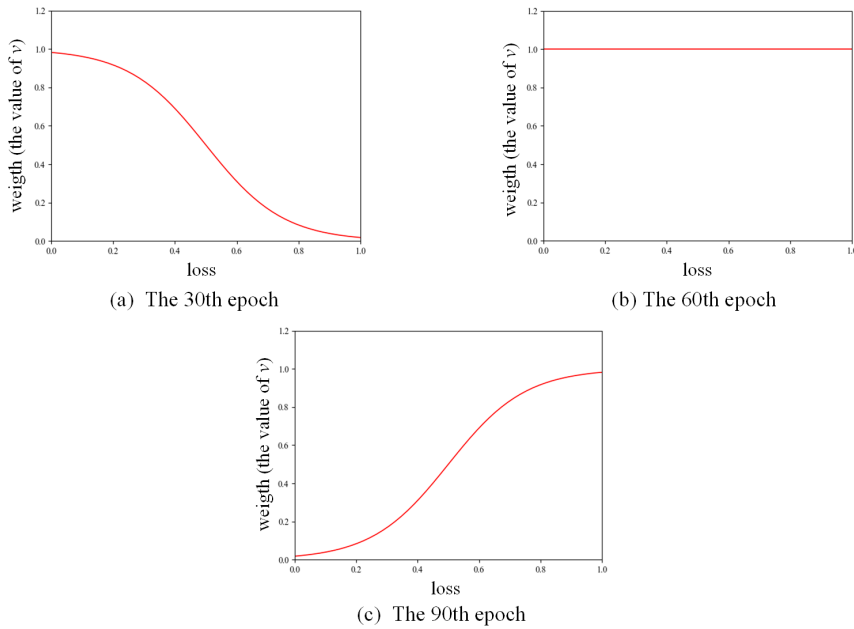


Figure 6: Curves of MSPL when τ is set to 0.5, α is set to 8, and T is set to 120.

- **BiLSTM–CRF (Glove).** Word embedding for each word is obtained through the Glove (Pennington et al., 2014), and then BiLSTM and CRF are used to obtain the label sequence of every word.
- **CNN–BiLSTM–CRF (Glove).** Character embedding for each word is obtained through the CNN network. Word embedding for each word is obtained through the Glove. The above-mentioned two embeddings are co-created to be fed into BiLSTM and CRF to obtain the label sequence.
- **CNN–BiLSTM–SA–CRF (Glove).** The difference between this model and CNN–BiLSTM–CRF is that SA is used to calculate the association among words in this model.
- **CNN–BiLSTM–SA–CRF (BERT).** A pretrained model, BERT (Devlin et al., 2019), is used to obtain the word embedding for each word.
- **CNN–BiLSTM–SA–CRF (SciBERT).** A pretrained model, SciBERT (Beltagy et al., 2019), is used to obtain the word embedding for each word.

Table 3

Evaluation results of different NER models.

Model	Original Sentences			Lowercase Sentences		
	Precision	Recall	F1	Precision	Recall	F1
BiLSTM-CRF (Glove)	0.7854	0.7753	0.7803	0.7644	0.6991	0.7303
CNN-BiLSTM-CRF (Glove)	0.7913	0.7829	0.7871	0.7798	0.7055	0.7408
CNN-BiLSTM-SA-CRF (Glove)	0.8044	0.7868	0.7955	0.7845	0.7651	0.7747
CNN-BiLSTM-SA-CRF (BERT)	0.8078	0.7935	0.8006	0.7974	0.7821	0.7897
CNN-BiLSTM-SA-CRF (SciBERT)	0.8116	0.7989	0.8052	0.8023	0.7894	0.7958
CNN-BiLSTM-SA-CRF+SPL (SciBERT)	0.8165	0.8052	0.8108	0.8056	0.7951	0.8003
CNN-BiLSTM-SA-CRF+MSPL (SciBERT)	0.8197	0.8106	0.8151	0.8098	0.8012	0.8055

- **CNN-BiLSTM-SA-CRF+SPL (SciBERT).** The difference between this model and CNN-BiLSTM-SA-CRF (SciBERT) is that SPL is used to decide which samples to participate in training.
- **CNN-BiLSTM-SA-CRF+MSPL (SciBERT).** MSPL is used to decide which samples to participate in training.

(2) Experimental settings

In all methods, the maximum sentence length is 100, and the batch size is 16. The hidden dimension of BiLSTM is 200. The character-level CNN network uses five parallel 3D convolution-activation-max pooling. Each of the five convolutions uses ten 3D convolution kernels ($10 \times 1 \times 50$, $1 \times 2 \times 50$, $1 \times 3 \times 50$, $1 \times 4 \times 50$, and $1 \times 5 \times 50$). Finally, the results obtained by five convolutions are spliced to obtain a 50D character embedding for every word. The hidden dimension of SA is 400. The 300D Glove (Pennington et al., 2014) embedding is used. The configurations of BERT and SciBERT are in accordance with that of Devlin et al. (2019) and Beltagy et al. (2019), respectively. In SPL, the initial weight of K is set to 4 and is reduced by a factor 1.1 at each iteration. In MSPL, the values of t_1 and t_2 are set as $3T/8$ and $5T/8$, respectively, where T is the number of total epochs and is set to 120. The value of τ and α are set to 0.5 and 8, respectively.

(3) Evaluation results

The raw and the corresponding lowercase samples are used to train the model. During the test, the test samples (1040 sentences) and the corresponding 1040 lowercase samples are tested. The evaluation results of the NER models are shown in Table 3.

Table 3 illustrates that MSPL improves the performance of the model, and CNN-BiLSTM-SA-CRF+MSPL (SciBERT) outperforms the other NER models. Accordingly, CNN-BiLSTM-SA-CRF+MSPL (SciBERT) is used to extract methods, datasets, and metrics from large-scale AI literature. To further ensure the reliability of extraction, we manually compile a dictionary of common methods, datasets, and metrics. After the NER model is used to extract AI markers, the dictionary is used to supplement unextracted entities and filter wrong entities. After combining with the dictionary, the F1 of CNN-BiLSTM-SA-CRF+MSPL (SciBERT) is 0.8813, the recall is 0.8792, and the precision is 0.8834.

4.2.3. AI marker normalization

In a paper, a method/dataset/metric may have multiple expressions. Our normalization approach follows the study of Wang and Zhang (2020). We compile a dictionary for methods, datasets, and metrics to normalize them. Specifically, we recruit 30 graduate students engaged in AI research to manually summarize all names of a method/dataset/metric. Table 4 shows some illustrative examples in our compiled dictionaries.

4.3. Original paper tracing for AI markers

The original papers of the methods and dataset must be traced back to obtain the research trace of a method or dataset that has been cited by other literature.

Table 4

Some illustrative examples in our compiled dictionary.

Dictionary	Normalized name	Abbreviations and aliases
Method	Long Short-Term Memory (LSTM)	LSTM, LSTMs, LSTM based, LSTM-based, Long Short Term Memory, Long Short-Term Memory, Long-Short-Term Memory, Long-Short-Term-Memory
	Recurrent Neural Net-work (RNN)	RNN, RNNs, RNN based, RNN-based, Recurrent NN, Recurrent NNs, Recurrent Neural Network, Recurrent Neural Networks
	Support Vector Machine (SVM)	SVM, SVMs, SVM based, SVM-based, Support Vector Machine, Support Vector Machines, Support-Vector Machine, Support-Vector Machines
Dataset	CIFAR	CIFAR, CIAFR10, CIAFR-10, CIFAR 10, CIFAR100, CIFAR-100, CIFAR 100
	SST	SST, SST1, SST-1, SST2, SST-2, SST5, SST-5
	COCO	COCO, MSCOCO, MS COCO, MS-COCO, Microsoft COCO, Microsoft-COCO, COCO2014, COCO 2014, COCO-2014, MSCOCO2014, MSCOCO 2014, MSCOCO-2014, MS COCO 2014, MS COCO2014, MS-COCO 2014, MS-COCO2014, Microsoft COCO 2014 (The year can be replaced with other numbers.)
Metric	F Measure	F Measure, F-Measure, F measures, F-measures, F Score, F-Score, F Scores, F-Scores, F1, F-1, F1 measure, F1 measures, F1-measure, F1-measures, F-1 measure, F-1 measures, F1 score, F1 scores, F1-score, F1-scores
	Mean Absolute Error (MAE)	MAE, Mean absolute error, MAEs, Mean absolute errors
	Adjusted Rand Index (ARI)	ARI, Adjusted Rand Index

4.3.1. Tracing approach

When a method or dataset is cited in a paper, the references for the corresponding original papers are often attached to it. In tracing, the set of papers citing the AI marker is firstly recorded for each AI marker. The sentences where the AI marker appears are located for each paper in the set. In each sentence, the existence of references in one or two positions behind the AI marker is checked. If a reference is present, then it is recorded. Finally, the most cited paper corresponding to each AI marker in the recorded references is selected as the original paper. Although this approach may produce errors, with such a large amount of data, we believe that the approach is a feasible and effective way to balance accuracy and cost in solving the problem.

4.3.2. Evaluation results

Using the above-mentioned tracing approach, a total of 5197 methods proposed in CCF corpus and 4166 corresponding original papers are obtained. Moreover, 1296 datasets proposed in CCF corpus and 971 corresponding original papers are obtained. The results are manually checked and the accuracy is 95.16%.

4.4. Clustering of AI markers

Methods using the same datasets and metrics are likely to solve the same task, but they may not appear in the same paper. Nonetheless, these methods are of the same type and it is significant to cluster them. Accordingly, we combine the datasets and metrics through the co-occurrence relationship and then merge the combined {datasets, metrics} and methods through the co-occurrence relationship to obtain the co-occurrence matrix of {methods, {datasets, metrics}}. Given the high-dimensional sparseness of the co-occurrence matrix, Nonnegative Matrix Factorization

(NMF) (Alghamedy and Zhang, 2018; Lee and Seung, 2001) and spectral clustering (Ng et al., 2002) are used together to build dimensionality reduction and clustering algorithms, and 500 method clusters are obtained⁴.

5. Results

This section performs statistical analysis, propagation analysis, and method cluster analysis on the basis of the AI markers in the collected CCF corpus (2005–2019 AI papers).

5.1. Statistics of AI markers

A total of 172,165 machine learning method entities, 16,877 dataset entities, and 1572 metric entities are mined by extracting the AI markers in the CCF corpus. Only AI markers that are cited more than once are considered in the analysis.⁵

This section introduces the analysis of AI markers in terms of the countries and publication venues. The top 10 AI markers used every year are also described.

5.1.1. Analysis in terms of countries

The number of AI markers proposed by a country can partially reflect its AI research level. The number of methods and datasets proposed by each country in the CCF corpus from 2005 to 2019 is calculated (Figs. 7 and 8)⁶.

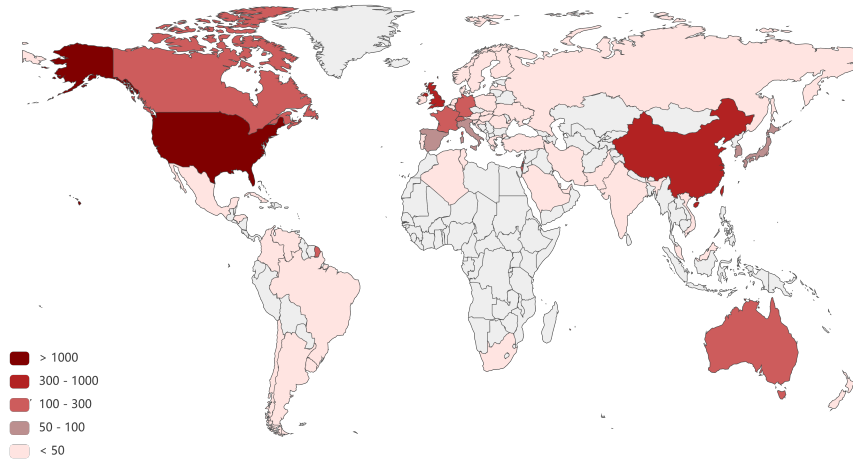


Figure 7: Distribution of the number of methods in different countries.

Fig. 7 demonstrates that the top three countries according to method quantities are the United States, China, and the United Kingdom, followed by Germany, France, Canada, Singapore, Australia and so on. In Fig. 8, the top three countries according to dataset quantities are the United States, China, and the United Kingdom, followed by Germany, Switzerland, Canada, France, Singapore, Israel and so on. The United States, China, and the United Kingdom are relatively active countries in the field of AI.

The proposal rates of the methods and datasets for each country are calculated to reduce the effect of the number of papers published in each country.

The proposal rate MR_c of the methods of country c and the proposal rate DR_c of the datasets of country c are calculated using Eqs. (9) and (10).

$$MR_c = \frac{|M_c|}{|L_c|}, \quad (9)$$

⁴The number of clusters represents the experimental parameters. It was verified by manually checking the clusters that the results are reasonable when the number of clusters are the above values.

⁵Considering the accuracy of the extracted entities, only entities with more than once in all papers are used for analysis.

⁶The country that proposed a specific AI marker is the country of the first author's institution of the original paper corresponding to the AI marker.

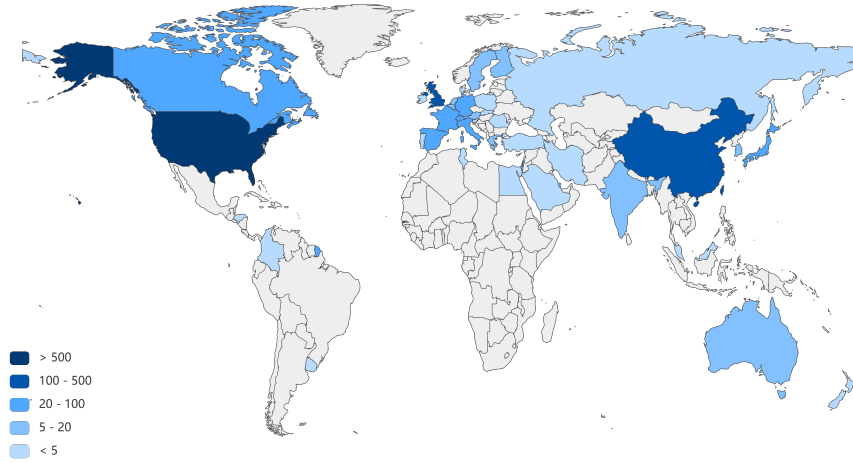


Figure 8: Distribution of the number of datasets in different countries.

$$DR_c = \frac{|D_c|}{|L_c|}, \quad (10)$$

where M_c is the set of methods proposed by country c , D_c is the set of datasets proposed by country c , and L_c is the set of papers proposed by country c .

The proposal rates of the methods of the top 10 countries in terms of the number of proposed methods and those of the datasets of the top 10 countries with regard to the number of proposed datasets are obtained on the basis of Eqs. (9) and (10). The results are shown in Fig. 9. The proposal rate of the methods of the United States ranked first. The proposal rate of Switzerland is the highest, which reflects that Switzerland attaches great importance to AI datasets.

5.1.2. Analysis in terms of the publication venues

To measure the quality of publication venues, the average usage of the entity proposed by each journal is calculated using Eq. (11).

$$U_v = \frac{\sum_{m \in M_v} C_m}{|M_v|} + \frac{\sum_{d \in D_v} C_d}{|D_v|}, \quad (11)$$

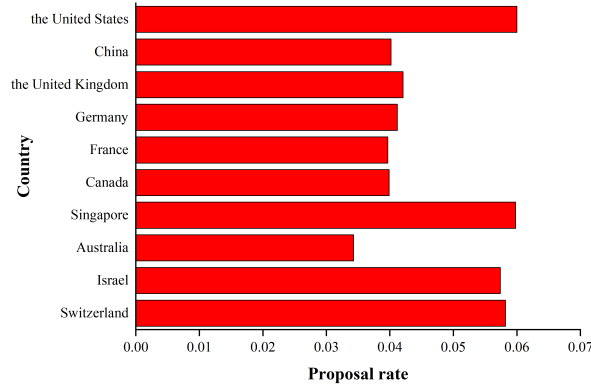
where M_v is the set of methods proposed by publication venue v , D_v is the set of datasets proposed by publication venue v , C_m is the number of citations of the original paper of the method m , and D_m is the number of citations of the original paper of the dataset d .

The top 10 publication venues in terms of average usage of the entity proposed are shown in Fig. 10. IJCV is ranked first. Furthermore, numerous publication venues are related to the computer vision (CV) field in the top 10, such as IJCV, TPAMI, CVPR, ICCV, and ECCV, indicating that entities in the CV field are more likely to be used by other papers. Among the top 10 publication venues, seven publication venues belong to Tier-A in CCF, which reflects that the quality of most papers in the publication venues of Tier-A is indeed higher than those of papers in Tier-B and Tier-C.

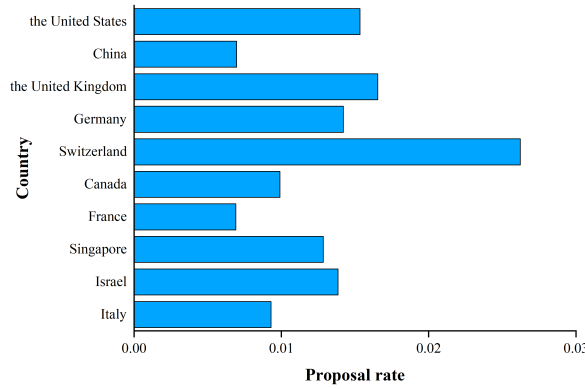
5.1.3. Annual top 10 AI markers

This section analyses the number of methods and datasets used every year from 2005 to 2019⁷. The number of methods used every year from 2005 to 2019 is counted. The top 10 popular methods used every year are shown in

⁷The methods/datasets used every year not only include the methods/datasets really used in the papers but also include the methods/datasets that are purely mentioned in the methodology and the experiment chapters because all these entities are important to the papers.



(a) Proposal rates of the methods of the top 10 countries listed in Fig. 7.



(b) Proposal rates of the datasets of the top 10 countries listed in Fig. 8.

Figure 9: Proposal rates of the AI markers of the top 10 countries listed in Figs. 7 and 8. The number of AI markers proposed by the countries decreased from top to bottom.

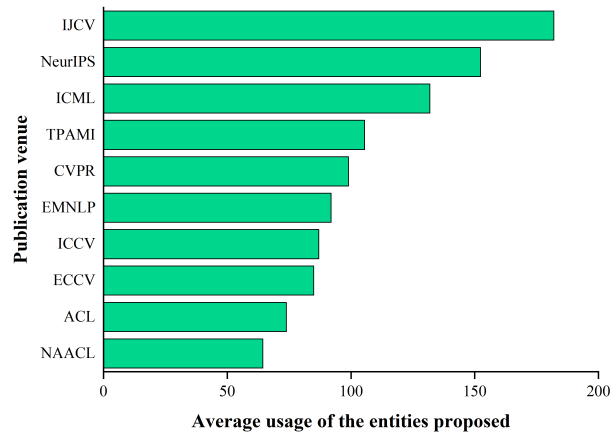


Figure 10: Top 10 publication venues in terms of average usage of the entity proposed.

Fig. 11. In Fig. 11, SVM, a traditional machine learning method, is widely used every year, and ranked first from 2005 to 2016. The proportion of SVM usage decreased after 2016 with the rapid development of deep learning. The deep learning method, CNN, is most widely used after 2016. Deep learning gains popularity in the AI field. One reason is that deep learning algorithms can be trained in an end-to-end manner without manual feature extraction. Furthermore, the deep learning algorithms can better learn the features with large-scale data and achieve excellent performance.

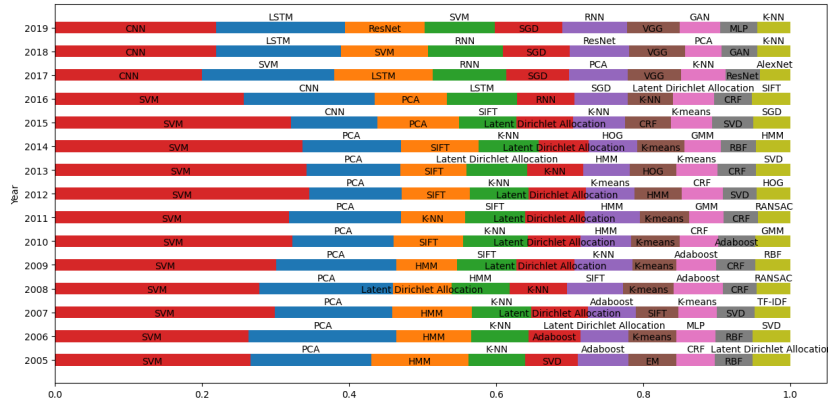


Figure 11: Top 10 methods used every year.

The number of datasets used every year is counted. The top 10 datasets used every year are shown in Fig. 12. MNIST ranked first from 2005 to 2014. ImageNet ranked first after 2014. The usage of ImageNet has increased over time. One reason is that ImageNet is large in scale and diversity, and facilitates the development of deep learning. The KITTI dataset (Geiger et al., 2012) is mainly used in the field of autonomous driving. The proportion of this dataset in the top 10 datasets gradually increased from 2017 to 2019, indicating that autonomous driving is growing in popularity. Many of the top 10 datasets are used in the face recognition field, such as Caltech, Yale, CMU PIE, and CelebA. The dataset proportion of face recognition in the top 10 datasets used every year is counted (Table 5). It can be seen that face recognition has always been a popular research direction.

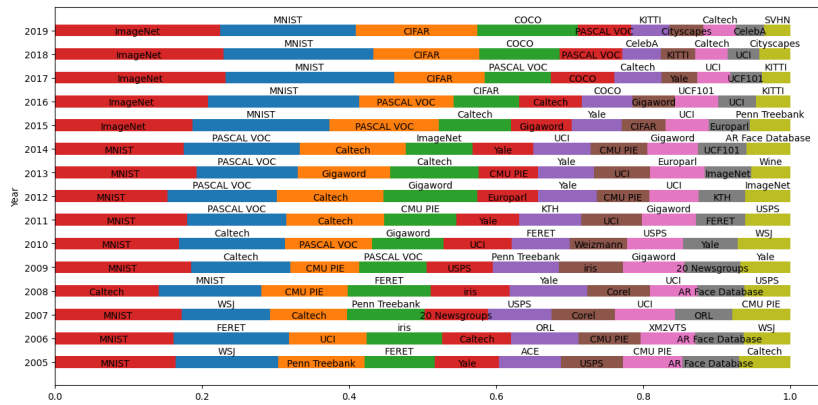


Figure 12: Top 10 datasets used every year.

Table 5

Proportion of datasets of face recognition in the top 10 datasets used every year.

Year	Face recognition	Proportion
2005	FERET, Yale, CMU PIE, AR Face Database, and Caltech	37.23%
2006	FERET, Caltech, ORL, CMU PIE, XM2VTS, and AR Face Database	56.67%
2007	Caltech, ORL, and CMU PIE	26.18%
2008	Caltech, CMU PIE, FERET, Yale, and AR Face Database	54.15%
2009	Caltech, CMU PIE, and Yale	29.62%
2010	Caltech, FERET, and Yale	29.79%
2011	Caltech, CMU PIE and Yale, and FERET	38.38%
2012	Caltech, Yale, and CMU PIE	29.61%
2013	Caltech, CMU PIE, and Yale	27.74%
2014	Caltech, Yale, CMU PIE, and AR Face Database	36.34%
2015	Caltech and Yale	16.50%
2016	Caltech	8.49%
2017	Caltech and Yale	11.24%
2018	CelebA and Caltech	9.64%
2019	Caltech	4.36%

5.2. Propagation of methods and datasets

This section analyses the propagation of methods among countries⁸ and the propagation of datasets.

5.2.1. Propagation of methods

Variable M_c is the set of all the methods proposed by country c , and $m \in M_c$. The propagation degree of the methods from country c to country c' in the time period from year y to year $y + \Delta y$ is calculated using Eq. (12).

$$PD_{c,c'}(\Delta y|y) = \sum_{m \in M_c} \left| LE_{c'}^m(\Delta y|y) \right| + \left| LM_{c'}^m(\Delta y|y) \right|, \quad (12)$$

where $LE_{c'}^m(\Delta y|y)$ is the set of papers in country c' , which cites method m in the experiment chapter in the time period from year y to year $y + \Delta y$, $LM_{c'}^m(\Delta y|y)$ is the set of papers in country c' , which cites method m in the methodology chapter in the time period from year y to year $y + \Delta y$, $y \in \{2005, 2006, \dots, 2019\}$, and $\Delta y \in \{0, 1, 2, \dots, 14\}$.

Based on Eq. (12), the propagation degrees of the methods among countries from 2005 to 2009, from 2010 to 2014 and from 2015 to 2019 are calculated. The top 10 propagation degrees among countries at each stage are shown in Fig. 13.

The methods are mainly propagated from the United States, France, and the United Kingdom to other countries from 2005 to 2009. The propagation degree of the methods proposed by China gradually increased from 2010 to 2014. Furthermore, the propagation degree of the methods proposed by China to the United States ranked second place from 2015 to 2019, indicating the rapid development of AI in China in recent years. The propagation degree of the methods proposed by the United States has been the first from 2005 to 2019.

5.2.2. Propagation of datasets

The propagation rate of the datasets proposed in year y in the time period from year y to year $y + \Delta y$ is calculated using Eq. (13).

$$PR(\Delta y|y) = \frac{\sum_{d \in D_y} |U_d(\Delta y|y)|}{|D_y|}, \quad (13)$$

⁸Given that the propagation analysis needs to be traced to the original papers of the methods, the methods and the corresponding original papers obtained in Section 4.3 are only used for analysis in this section.

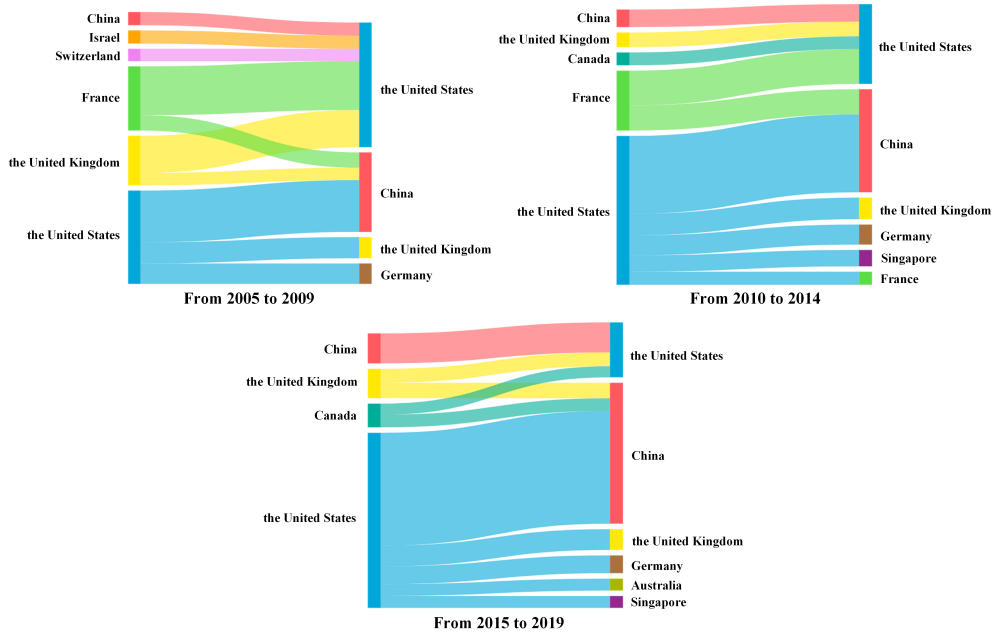


Figure 13: Top 10 propagation degrees of methods among countries from 2005 to 2019. The usage from the same country is excluded, and the propagation direction is from left to right.

where D_y is the set of all the datasets proposed in year y , $U_d(\Delta y|y)$ is the set of the papers that used the dataset d in the time period from year y to year $y + \Delta y$, and $\Delta y \in \{0, 1, 2\}$.

Based on Eq. (13), the propagation rates of the datasets are obtained and shown in Fig. 14. The aforementioned figure demonstrates that the propagation rate gradually increases over time, indicating that researchers pay more and more attention to datasets.

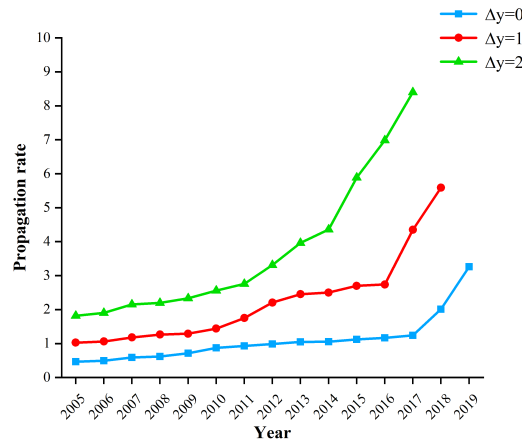
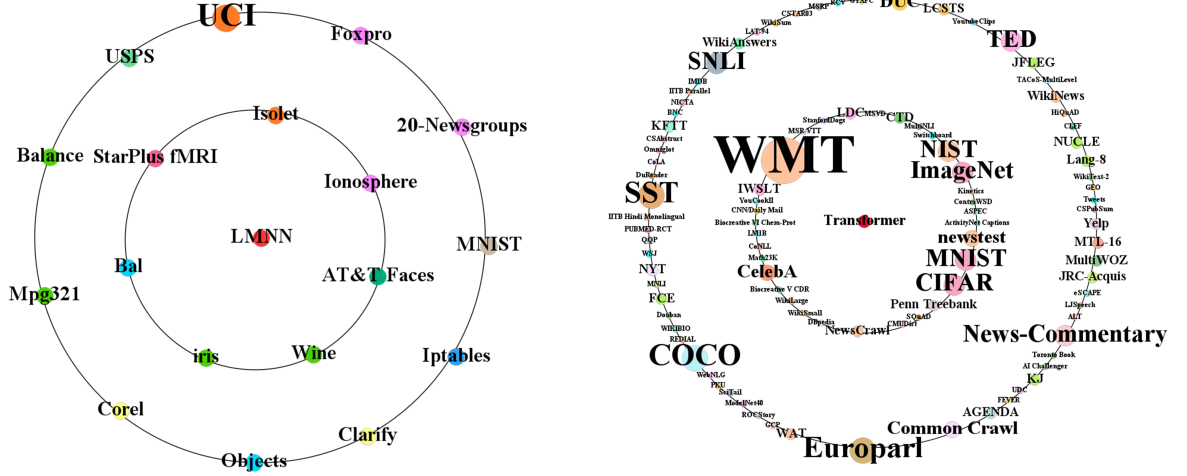


Figure 14: Propagation rates of datasets.

We also compare the application of the methods on the datasets within a few years after they were proposed. The large margin nearest neighbour (LMNN) and Transformer methods proposed in 2005 and 2018, respectively, are considered in the case study. The results are shown in Fig. 15.

Fig. 15 demonstrates that after Transformer was proposed, it was quickly applied to various datasets in 2018 and 2019. LMNN was proposed in 2005. However, the number of datasets applying LMNN was small in 2006 and 2007.

Moreover, the number and types of applications of Transformer are far more than those of LMNN. This notion reflects that the propagation has been increased rapidly in recent years. One reason lies in the increasing involvement of researchers in the field due to the rapid development of deep learning. The depth and breadth of the related studies have been continuously improved, and multiple datasets have been used to fully validate the performance of the algorithms. Additionally, deep learning algorithms have become the primary approaches in various AI subfields in recent years, so that the advances in one subfield can be applied to other subfields quickly.



(a) Application of LMNN to the datasets in 2006 (inner circle) and 2007 (b) Application of Transformer to the datasets in 2018 (inner circle) and 2019

Figure 15: Application of different methods on the datasets. The red point in the centre indicates the method. The inner and outer circles are composed of many dataset points. In the dataset points, the size of the point indicates the number of the dataset applied to the method.

5.3. Analysis of method clusters

5.3.1. Roadmap generation for method clusters

A method roadmap describes the evolution of different yet highly correlated methods (Zha et al., 2019). In the method clusters obtained by the clustering algorithm, each method cluster is composed of different yet highly correlated methods. In the cluster, if a time-based method evolution map can be built, and the dataset information can be added, then it will provide enlightening information for related research.

The roadmap generation procedure for a method cluster (the output of the algorithm in Section 4.4) proposed in the study is as follows:

1) The information of all the methods in the cluster is obtained, including the time of proposal, the chapters where the method locates in its original paper and the associated datasets used in its original paper⁹.

2) For each method M_i in the cluster, the other methods $\{M_1, \dots, M_n\}$ mentioned in the experiment chapter of the original paper of M_i are found. Path $M_i \rightarrow M_j$ from M_i to M_j is constructed, and $M_j \in \{M_1, \dots, M_n\}$. The edge between M_i and M_j represents the datasets used for the comparison between M_i and M_j .

3) Continuous paths are combined to obtain the roadmap of the methods in the same method cluster. For example, if $(M_1 \rightarrow M_2)$, $(M_2 \rightarrow M_3)$ and $(M_1 \rightarrow M_3)$ exist, then only $(M_1 \rightarrow M_2)$ and $(M_2 \rightarrow M_3)$ are kept.

The construction of our roadmaps differs from the approach in Zha et al. (2019) in two points: 1) The datasets are added in our roadmap construction, which provides additional information; 2) the methods are obtained by large-scale literature mining in our roadmap, and numerous roadmaps can be simultaneously obtained.

⁹If the original paper is not traced back in Section 4.3, then the earliest paper cited by this method will be used as the original paper.

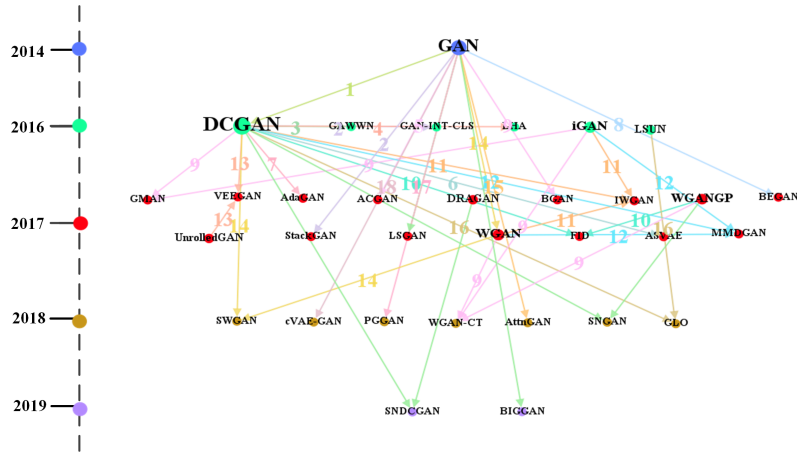


Figure 17: Roadmap of the methods in the 'GAN' cluster. The numbers in the figure indicate the datasets used when comparing M_i and M_j in path $M_i \rightarrow M_j$. The coloured dots in the figure indicate the years, their sizes denote the size of the out-degrees, and the coloured lines signify the datasets represented by the numbers. The correspondence between the numbers and the datasets is as follows: 1: MNIST, SVHN, and CelebA; 2: CUB and Oxford Flower; 3: CUB, MPII Human Pose, Caltech, and MHP; 4: ImageNet and SVHN; 5: ImageNet; 6: MNIST, CIFAR, and ImageNet; 7: MNIST; 8: CelebA; 9: MNIST, CIFAR, and SVHN; 10: SVHN, CIFAR, CelebA, and LSUN Bedroom; 11: LSUN Bedroom; 12: MNIST, CIFAR, CelebA, and LSUN Bedroom; 13: MNIST and CIFAR; 14: CIFAR and LSUN Bedroom; 15: ImageNet and COCO; 16: MNIST, SVHN, CelebA, and LSUN Bedroom; 17: CelebA and LSUN Bedroom; 18: Chinese poem.

paper tracing results. The results show that the United States, China, and the United Kingdom are relatively active countries in the field of AI. The propagation rate of the datasets gradually increases over time. The methods proposed by China in recent years have an increasing influence on other countries. Finally, the datasets and metrics are combined, and the combined results are combined with methods to get the co-occurrence matrix, which is used for clustering. The roadmaps of the methods are drawn on the basis of the method clusters and associated datasets to study the evolution of the methods in the same cluster.

In summary, we construct a large-scale literature dataset, which can be valuable for the community. Then, the development in the field of AI is explored from the perspectives of methods, datasets, and metrics, which can benefit help many researchers, such as beginners in AI research, in quickly and accurately finding research information and understand research trends. In addition, the MSPL strategy is proposed to address the negative influence of hard and noisy samples in training. The MSPL strategy can also be used in other tasks, such as classification tasks.

Nevertheless, this paper still has some limitations. Firstly, the year-range of papers we collected is limited. We only collected papers from 2005 to 2019. In the future, we will collect more papers to more fully reflect the development of the AI field. Secondly, during the tracing the original paper tracing for a specific AI marker, we only consider citation information. However, the approach may cause some errors. For example, sometimes, a paper cited by an author about a specific algorithm may be the most famous paper for the algorithm applied to a certain task, rather than the original paper of the algorithm. In this work, with such a large amount of data, we believe that the approach is a feasible and effective way to balance accuracy and cost in solving the problem. In the future, we will explore a better way to solve this problem. Finally, we do not study the development of different AI subfields, such as the similarities and differences between the NLP field and the CV field, which is also a very important issue. Given that many conferences and journals publish papers in NLP and CV fields, how to divide the fields of papers is a challenging problem. We will explore this issue in the future.

In addition to the above-mentioned future research directions, we will also design a better model structure to extract entities more accurately. Furthermore, we will consider how to utilize recommendation algorithms to recommend algorithms or datasets to scholars. Finally, we will consider extracting more kinds of entities from AI literature, such as optimization methods, to achieve a more comprehensive presentation of the AI field.

References

- Adhikari, A., Ram, A., Tang, R., Lin, J., 2019. Rethinking complex neural network architectures for document classification, 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), pp. 4046–4051.
- Alghamedy, F., Zhang, J., 2018. Enhance nmf-based recommendation systems with social information imputation. *Computer Science & Information Technology (CS & IT). AIRCC*, 37–54.
- Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620.
- Chen, J., Liu, Q., Lin, H., Han, X., Sun, L., 2022. Few-shot named entity recognition with self-describing networks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5711–5722.
- Cheng, Q., Li, P., Zhang, G., Lu, W., 2021. Recognition of lexical functions in academic texts: Problem method extraction based on title generation strategy and attention mechanism. *Journal of the China Society for Scientific and Technical Information* 40, 43–52.
- Cho, H.C., Okazaki, N., Miwa, M., Tsujii, J., 2013. Named entity recognition with multiple segment representations. *Information Processing & Management* 49, 954–965.
- Constantin, A., Pettifer, S., Voronkov, A., 2013. Pdfx: fully-automated pdf-to-xml conversion of scientific literature, in: Proceedings of the 2013 ACM symposium on Document engineering, pp. 177–180.
- Dai, Z., Fei, H., Li, P., 2019. Coreference aware representation learning for neural named entity recognition., in: IJCAI, pp. 4946–4953.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255.
- Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K., 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* 51, 32–49.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. 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, NAACL-HLT 2019, Volume 1 (Long and Short Papers), pp. 4171–4186.
- Ding, R., Wang, Y., Zhang, C., 2019. Investigating citation of algorithm in full-text of academic articles in nlp domain: A preliminary study, in: Proceedings of the 17th international conference on scientometrics and informetrics (ISSI 2019), Rome, Italy, pp. 2726–2728.
- Ding, Y., Liu, X., Guo, C., Cronin, B., 2013a. The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics* 7, 583–592.
- Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., Chambers, T., 2013b. Entitymetrics: Measuring the impact of entities. *PloS one* 8, e71416.
- Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M., 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, in: Proceedings of NAACL-HLT, pp. 2368–2378.
- Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C.D., Stamatiopoulos, P., 2000. Rule-based named entity recognition for greek financial texts, in: Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000), pp. 75–78.
- Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361.
- Heddes, J., Meerdink, P., Pieters, M., Marx, M., 2021. The automatic detection of dataset names in scientific articles. *Data* 6, 84.
- Heffernan, K., Teufel, S., 2018. Identifying problems and solutions in scientific text. *Scientometrics* 116, 1367–1382.
- Hong, Y., Hwang, U., Yoo, J., Yoon, S., 2019. How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)* 52, 1–43.
- Huang, S., Wan, X., 2013. Akminer: Domain-specific knowledge graph mining from academic literatures, in: International Conference on Web Information Systems Engineering, pp. 241–255.
- Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y., 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* 33, 494–514.
- Ji, Y., Zhang, Q., Shen, S., Wang, D., Huang, S., 2019. Research on functional structure identification of academic text based on deep learning, in: In Proceedings of 17th International Conference of the International-Society-for-Scientometrics-and-Informetrics (ISSI), Vol II, pp. 2712–2713.
- Jie, Z., Xie, P., Lu, W., Ding, R., Li, L., 2019. Better modeling of incomplete annotations for named entity recognition, 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), pp. 729–734.
- Kim, J., Le, D.X., Thoma, G.R., 2000. Automated labeling in document images, in: Document Recognition and Retrieval VIII, pp. 111–122.
- Kondo, T., Nanba, H., Takezawa, T., Okumura, M., 2009. Technical trend analysis by analyzing research papers' titles, in: Language and Technology Conference, pp. 512–521.
- Kumar, M., Packer, B., Koller, D., 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems* 23.
- Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 2278–2324.
- Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization, in: Advances in neural information processing systems, pp. 556–562.
- Lei, Z., Wang, D., 2019. Model entity extraction in academic full text based on deep learning, in: Proceedings of the 17th international conference on scientometrics and informetrics (ISSI), Vol II, pp. 2732–2733.
- Li, F., Lin, Z., Zhang, M., Ji, D., 2021a. A span-based model for joint overlapped and discontinuous named entity recognition, 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. 4814–4828.

- Li, P., Liu, Q., Cheng, Q., Lu, W., 2021b. Data set entity recognition based on distant supervision. *The Electronic Library* .
- Li, X., Cheng, Q., Lu, W., 2017. CS-LAS: A scientific literature retrieval and analysis system based on term function recognition (TFR), in: *In Proceedings of the 16th International Conference on Scientometrics and Informetrics, ISSI*, pp. 1346–1356.
- Li, Y., Shi, S., et al., 2020. Empirical analysis of unlabeled entity problem in named entity recognition, in: *International Conference on Learning Representations*.
- Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.
- Lu, W., Huang, Y., Bu, Y., Cheng, Q., 2018. Functional structure identification of scientific documents in computer science. *Scientometrics* 115, 463–486.
- 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, 1320–1329.
- Lu, W., Li, X., Liu, Z., Cheng, Q., 2019. How do author-selected keywords function semantically in scientific manuscripts? *Knowledge Organization: KO* 46, 403.
- Luo, Y., Xiao, F., Zhao, H., 2020. Hierarchical contextualized representation for named entity recognition, in: *Proceedings of the AAAI conference on artificial intelligence*, pp. 8441–8448.
- Ma, B., Zhang, C., Wang, Y., 2021. Exploring significant characteristics and models for classification of structure function of academic documents. *Data and Information Management* 5, 65–74.
- Ma, B., Zhang, C., Wang, Y., Deng, S., 2022a. Enhancing identification of structure function of academic articles using contextual information. *Scientometrics* 127, 885–925.
- Ma, Y., Liu, J., Lu, W., Cheng, Q., 2022b. Beyond tasks, methods, and metrics: extracting metrics-driven mechanism from the abstracts of ai articles, in: *EEKE@ JCDL*.
- Mollá, D., Van Zaanen, M., Smith, D., 2006. Named entity recognition for question answering, in: *Proceedings of the Australasian language technology workshop 2006*, pp. 51–58.
- Nanba, H., Kondo, T., Takezawa, T., 2010. Automatic creation of a technical trend map from research papers and patents, in: *Proceedings of the 3rd international workshop on Patent information retrieval*, pp. 11–16.
- Ng, A.Y., Jordan, M.I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm, in: *Advances in neural information processing systems*, pp. 849–856.
- Pan, X., Yan, E., Wang, Q., Hua, W., 2015. Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics* 9, 860–871.
- Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.
- Ratinov, L., Roth, D., 2009. Design challenges and misconceptions in named entity recognition, in: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 147–155.
- Riaz, K., 2010. Rule-based named entity recognition in urdu, in: *Proceedings of the 2010 named entities workshop*, pp. 126–135.
- Ruben, S., Randall, M., Kamen, M., Hyde, J.L., 1941. Heavy oxygen (o18) as a tracer in the study of photosynthesis. *Journal of the American Chemical Society* 63, 877–879.
- Shang, F., Ran, C., 2022. An entity recognition model based on deep learning fusion of text feature. *Information Processing & Management* 59, 102841.
- Siekmeier, A., Lee, W., Kwon, H., Lee, J.H., 2021. Tag assisted neural machine translation of film subtitles, in: *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pp. 255–262.
- Tsai, C.T., Kundu, G., Roth, D., 2013. Concept-based analysis of scientific literature, in: *Proceedings of the 22nd ACM international conference on information & knowledge management*, pp. 1733–1738.
- Tuarob, S., Mitra, P., Giles, C.L., 2015. A hybrid approach to discover semantic hierarchical sections in scholarly documents, in: *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 1081–1085.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: *Advances in neural information processing systems*, pp. 5998–6008.
- Wan, J., Ru, D., Zhang, W., Yu, Y., 2022. Nested named entity recognition with span-level graphs, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 892–903.
- Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., Tu, K., 2021. Improving named entity recognition by external context retrieving and cooperative learning, 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. 1800–1812.
- Wang, Y., Zhang, C., 2018. Using full-text of research articles to analyze academic impact of algorithms, in: *International Conference on Information*, pp. 395–401.
- Wang, Y., Zhang, C., 2020. 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, 101091.
- Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., Han, J., 2019. Crossweigh: Training named entity tagger from imperfect annotations, in: *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pp. 5154–5163.
- Zha, H., Chen, W., Li, K., Yan, X., 2019. Mining algorithm roadmap in scientific publications, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1083–1092.
- Zhang, C., Ding, R., Wang, Y., 2021a. Algorithms mention in full-text content of article from nlp domain: Comparative analysis between english and chinese. *Data Science and Informetrics* 1, 19–33.
- Zhang, C., Xie, Y., Song, Y., 2021b. Association analysis of fine-grained knowledge entities in academic texts. *Library Tribune* 41, 12–20.

- Zhang, H., Ma, S., Zhang, C., 2019. Using full-text of academic articles to find software clusters, in: ISSI, pp. 2776–2777.
- Zhang, Y., Zhao, R., Wang, Y., Chen, H., Mahmood, A., Zaib, M., Zhang, W.E., Sheng, Q.Z., 2022. Towards employing native information in citation function classification. *Scientometrics* , 1–21.
- Zhao, H., Tian, X., He, L., Li, Y., Pu, W., Liu, Q., Tang, J., Wu, J., Cheng, X., Liu, Y., et al., 2018a. Apj+ vessels drive tumor growth and represent a tractable therapeutic target. *Cell reports* 25, 1241–1254.
- Zhao, M., Yan, E., Li, K., 2018b. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology* 69, 32–46.
- Zhao, N., Kamijo, K., Fox, P.D., Oda, H., Morisaki, T., Sato, Y., Kimura, H., Stasevich, T.J., 2019. A genetically encoded probe for imaging nascent and mature ha-tagged proteins in vivo. *Nature communications* 10, 1–16.
- Zhu, E., Li, J., 2022. Boundary smoothing for named entity recognition, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7096–7108.

Table A1

Information of collected papers.

No.	Publication Venue (Abbreviation)	Website	Number
1	AAAI Conference on Artificial Intelligence (AAAI)	http://dblp.uni-trier.de/db/conf/aaai/	8295
2	Annual Conference on Neural Information Processing Systems (NeurIPS)	http://dblp.uni-trier.de/db/conf/nips/	5604
3	Annual Meeting of the Association for Computational Linguistics (ACL)	http://dblp.uni-trier.de/db/conf/acl/	9207
4	IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	http://dblp.uni-trier.de/db/conf/cvpr/	11049
5	International Conference on Computer Vision (ICCV)	http://dblp.uni-trier.de/db/conf/iccv/	4887
6	International Conference on Machine Learning (ICML)	http://dblp.uni-trier.de/db/conf/icml/	4198
7	International Joint Conference on Artificial Intelligence (IJCAI)	http://dblp.uni-trier.de/db/conf/ijcai/	4228
8	Annual Conference on Computational Learning Theory (COLT)	http://dblp.uni-trier.de/db/conf/colt/	743
9	Conference on Empirical Methods in Natural Language Processing (EMNLP)	http://dblp.uni-trier.de/db/conf/emnlp/	3927
10	European Conference on Artificial Intelligence (ECAI)	http://dblp.uni-trier.de/db/conf/ecai/	2582
11	European Conference on Computer Vision (ECCV)	http://dblp.uni-trier.de/db/conf/eccv/	3657
12	International Conference on Case-Based Reasoning and Development (IC-CBR)	http://dblp.uni-trier.de/db/conf/iccbr/	515
13	International Conference on Computational Linguistics (COLING)	http://dblp.uni-trier.de/db/conf/coling/	2365
14	International Conference on Principles of Knowledge Representation and Reasoning (KR)	http://dblp.uni-trier.de/db/conf/kr/	509
15	International Conference on Uncertainty in Artificial Intelligence (UAI)	http://dblp.uni-trier.de/db/conf/uai/	1320
16	Artificial Intelligence and Statistics (AISTATS)	http://dblp.uni-trier.de/db/conf/aistats/	1822
17	Asian Conference on Computer Vision (ACCV)	http://dblp.uni-trier.de/db/conf/accv/	1857
18	Asian Conference on Machine Learning (ACML)	http://dblp.uni-trier.de/db/conf/acml/	401
19	British Machine Vision Conference (BMVC)	http://dblp.uni-trier.de/db/conf/bmvc/	1692
20	CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)	https://dblp.uni-trier.de/db/conf/nlpcc/	25
21	Conference on Computational Natural Language Learning (CoNLL)	http://dblp.uni-trier.de/db/conf/conll/	888
22	IEEE International Conference on Tools with Artificial Intelligence (ICTAI)	http://dblp.uni-trier.de/db/conf/ictai/	1987
23	International Conference on Algorithmic Learning Theory (ALT)	http://dblp.uni-trier.de/db/conf/alt/	329
24	International Conference on Artificial Neural Networks (ICANN)	http://dblp.uni-trier.de/db/conf/icann/	1652
25	International Conference on Automatic Face and Gesture Recognition (FG)	http://dblp.uni-trier.de/db/conf/fg/	900
26	International Conference on Document Analysis and Recognition (ICDAR)	http://dblp.uni-trier.de/db/conf/icdar/	2012
27	International Conference on Inductive Logic Programming (ILP)	http://dblp.uni-trier.de/db/conf/ilp/	354
28	International conference on Knowledge Science, Engineering and Management (KSEM)	http://dblp.uni-trier.de/db/conf/ksem/	585
29	International Conference on Neural Information Processing (ICONIP)	http://dblp.uni-trier.de/db/conf/iconip/	3547
30	International Conference on Pattern Recognition (ICPR)	http://dblp.uni-trier.de/db/conf/icpr/	6298
31	International Joint Conference on Neural Networks (IJCNN)	http://dblp.uni-trier.de/db/conf/ijcnn/	8044
32	Pacific Rim International Conference on Artificial Intelligence (PRICAI)	http://dblp.uni-trier.de/db/conf/pricai/	671
33	The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)	http://dblp.uni-trier.de/db/conf/naacl/	3664
34	IEEE Trans on Pattern Analysis and Machine Intelligence (TPAMI)	http://dblp.uni-trier.de/db/journals/pami/	2889
35	International Journal of Computer Vision (IJCV)	http://dblp.uni-trier.de/db/journals/ijcv/	1292
36	Journal of Machine Learning Research (JMLR)	http://dblp.uni-trier.de/db/journals/jmlr/	824
37	Autonomous Agents and Multi-Agent Systems (AAMAS)	http://dblp.uni-trier.de/db/journals/aamas/	441
38	IEEE Transactions on Audio, Speech, and Language Processing (TASLP)	http://dblp.uni-trier.de/db/journals/taslp/	1644
39	IEEE Transactions on Fuzzy Systems (TFS)	http://dblp.uni-trier.de/db/journals/tfs/	1459
40	Journal of Automated Reasoning	http://dblp.uni-trier.de/db/journals/jar/	459
41	Machine Learning	http://dblp.uni-trier.de/db/journals/ml/	860
42	Applied Intelligence	http://dblp.uni-trier.de/db/journals/apin/	1524
43	International Journal on Document Analysis and Recognition (IJ DAR)	http://dblp.uni-trier.de/db/journals/ijdar/	317
44	Machine Translation	http://dblp.uni-trier.de/db/journals/mt/	186
45	Machine Vision and Applications	http://dblp.uni-trier.de/db/journals/mva/	913
46	Natural Computing	http://dblp.uni-trier.de/db/journals/nc/	656
47	Neural Computing & Applications (NCA)	http://dblp.uni-trier.de/db/journals/nca/	3624
48	Neural Processing Letters (NPL)	http://dblp.uni-trier.de/db/journals/npl/	1082
49	Pattern Analysis and Applications (PAA)	http://dblp.uni-trier.de/db/journals/paa/	770
50	Soft Computing	http://dblp.uni-trier.de/db/journals/soco/	3692