Scientometrics (2023) 128:6335–6355
https://doi.org/10.1007/s11192-023-04858-4

Multi-task learning model for citation intent classification in scientific publications

Ruihua Qi[1,2] · Jia Wei[2] · Zhen Shao[2] · Zhengguang Li[1] · Heng Chen[1] · Yunhao Sun[1] · Shaohua Li[1]

Received: 7 September 2022 / Accepted: 13 October 2023 / Published online: 28 October 2023

Abstract
Citations play a significant role in the evaluation of scientific literature and researchers. Citation intent analysis is essential for academic literature understanding. Meanwhile, it is useful for enriching semantic information representation for the citation intent classification task because of the rapid growth of publicly accessible full-text literature. However, some useful information that is readily available in citation context and facilitates citation intent analysis has not been fully explored. Furthermore, some deep learning models may not be able to learn relevant features effectively due to insufficient training samples of citation intent analysis tasks. Multi-task learning aims to exploit useful information between multiple tasks to help improve learning performance and exhibits promising results on many natural language processing tasks. In this paper, we propose a joint semantic representation model, which consists of pretrained language models and heterogeneous features of citation intent texts. Considering the correlation between citation intents, citation section and citation worthiness classification tasks, we build a multi-task citation classification framework with soft parameter sharing constraint and construct independent models for multiple tasks to improve the performance of citation intent classification. The experimental results demonstrate that the heterogeneous features and the multi-task framework with soft parameter sharing constraint proposed in this paper enhance the overall citation intent classification performance.

Keywords Citation intent classification · Multi-task · Pretrained language model · Heterogeneous features

✉ Ruihua Qi
rhqi@dlufl.edu.cn

Zhengguang Li
lizhengguang2004@163.com

[1] School of Software, Dalian University of Foreign Languages, Dalian, Liaoning, People's Republic of China

[2] Research Center for Language Intelligence, Dalian University of Foreign Languages, Dalian, Liaoning, People's Republic of China

Introduction

Citation intent refers to the author's inner psychological activities when quoting documents, which reflects the reason and purpose of the citation. The goal of citation intent classification research is to study how to classify an author's citation purpose into a specific category. Citation intent classification analysis is crucial for understanding academic literature and identifying critical references, given citations' significant role in the evaluation of scientific research behavior. In addition, an in-depth study of citation intent classification is helpful for other bibliometric analysis tasks, such as citation importance analysis, academic impact evaluation, potential scholar relationship discovery, citation information retrieval, and so on (Garfield, 1972).

Early research on citation intent classification mainly relied on manual qualitative analysis of small sample data. For example, Hassan and Serenko (2019) point out that the author's citation selection processes are subjective and vary among researchers by their background. Accordingly, they develop a qualitative citation patterns typology to understand the motivations behind citations and illustrate how these citation patterns can be identified (Hassan & Serenko, 2019). However, when considering the generalization and scalability, manual qualitative methods can hardly cope with the rapidly growing accessible full-text literature. The development of natural language processing technology makes it possible to automatically extract semantic information on large-scale literature texts for citation intent analysis. Given that the existing citation features used in the citation intent classification task mainly include content, location, sentence, syntactic, structural features, and so on (Paice, 1990), it is impractical to extract sufficient semantic information by a set of predefined features to represent complex citation intents. Therefore, deep learning based models become widely used in citation intent analysis by representing the semantic information in citation contexts via pretrained language models. These models achieve good performance and outperform previous feature engineering based citation intent analysis methods (Andrade and Gonçalves 2020; Prester et al., 2021).

Despite achieving promising results, deep learning-based models for citation intent prediction still face several limitations that need to be addressed. One major challenge is the scarcity of large-scale annotated datasets. Training neural networks to accurately predict citation intents requires a significant amount of pre-annotated data, which is both expensive and time-consuming to obtain. Although some annotated citation intent datasets exist, they are typically small in size. To overcome this challenge, multi-task deep learning models have been proposed, enabling the sharing of bottom hidden layer parameters across auxiliary citation tasks (Cohan et al., 2019; Yousif et al., 2019). This approach allows the model to leverage existing annotated datasets from other citation-related tasks and reduces the risk of overfitting.

However, the hard parameter sharing mechanism, which assumes that all tasks share the same set of parameters, can limit the model's adaptability to task-specific features. The constraints imposed by the hard parameter sharing mechanism can be overly restrictive on auxiliary tasks, resulting in a sharp decline in performance when the tasks are not closely related (Ruder, 2017). Furthermore, hard parameter sharing makes it difficult for multi-task models to effectively represent the heterogeneous citation features originating from multiple sources. In summary, while deep learning-based models for citation intent prediction show great promise, they still face significant challenges, including the scarcity of large-scale annotated datasets and the limited adaptability to task-specific features.

This paper aims to address two issues faced by deep learning-based models for citation intent classification. The first issue is the scarcity of large-scale annotated datasets, while the second issue is the limited adaptability to task-specific features. To tackle these challenges, this study optimized the learning process of the citation intent classification task with auxiliary tasks, such as citation worthiness classification and citation section classification. By leveraging these auxiliary tasks, the model can learn from additional annotated data and enhance its ability to handle task-specific features.

The main contributions of this study are as follows:

- A novel multi-task deep learning framework is proposed for citation intent classification in this paper. The framework could effectively learn valuable information from annotated data of various scales and related tasks by employing a soft parameter sharing mechanism. This mechanism enhanced the model's adaptability to task-specific features while also benefiting from shared bottom hidden layer parameters. With this approach, the framework achieves improved performance by leveraging the strengths of both shared knowledge and task-specific adaptation.
- A joint citation intent representation model is proposed in this paper, which combines pretrained language models with various heterogeneous features. The purpose of this model is to complement the information that is not captured by the pretrained language models. As a result, it helped to enhance the overall performance of the proposed framework for citation intent classification.
- A series of experiments were conducted to analyze the impact of various factors on the performance of citation intent classification. These factors included the selection of auxiliary tasks, the utilization of different functional modules in the framework, and detailed per-category results. Our approach achieved state-of-the-art performance in citation intent classification tasks on two benchmark datasets, demonstrating its effectiveness in capturing implicit citation intent patterns.

The rest of this article is organized as follows. "Literature review" section introduces the background literature in detail on existing citation intent classification techniques. "Methodology" section describes our proposed joint representation model and multi-task deep learning framework for citation intent classification. "Experimentation and results" section presents the results of our experiments and compares them with a series of state-of-the-art methods on public citation analysis datasets. Finally, "Conclusion and future work" section presents our conclusions and indicates directions for future research.

Literature review

In previous studies, there are mainly three automatic citation intent classification methods, including citation analysis based on feature engineering and machine learning, citation intent analysis based on deep learning, and citation intent analysis based on multi-task learning.

Citation analysis based on feature engineering and machine learning

Citation analysis based on feature engineering involves two main steps: extracting citation features from citation texts and learning models using machine learning algorithms based

on these features. Over the years, researchers have explored the possibility of automatic citation analysis based on feature engineering, and various approaches have been proposed.

To analyze in-text citations, Teufel and Moens (2002) utilize sentential features, meta-discourse features, and Paice's (1990) features. They employ a supervised IBK classifier to summarize scientific articles. They also introduce a supervised machine learning framework that employs basic features, such as syntactic features like Part-of-Speech (POS), for citation function classification, rather than manual annotation (Teufel et al., 2006). Additionally, Dong and Schäfer (2011) as well as Jochim and Schütze (2012) use expression patterns of POS tags to capture syntactic information for analyzing citation intent. Teufel et al. (2006) discover a strong relationship between citation function and sentiment classification. The effectiveness of sentiment features in fine-grained citation analysis tasks has been demonstrated (Jochim & Schütze, 2012). Furthermore, the combination of TF-IDF features has also been proven to be beneficial in citation classification tasks (Andrade and Gonçalves 2020; Oesterling et al., 2021). Jurgens et al. (2018) combine pattern, topic, grammatical, and metadata features to identify citation intents using a Random Forest classifier. Tuarob et al. (2019) find that cue word features yield the best average F1 score, while structure, lexical-morphological-grammatical, sentiment, and venue features contribute to the recognition of certain intent classes.

In the analysis of paper-level citations, Valenzuela et al. (2015) employ SVM and Random Forests, considering citation location, citation count, author overlap, and abstract similarity to identify citation importance. In an extension of Valenzuela's work, Hassan et al. (2018) introduce a feature set for citation importance analysis, including demonstrative determiners, closest verbs, adjectives, adverbs, and more. Xu et al. (2013) devised nine network features, including whether the citation is self-citation, the frequency of the cited paper's mention, the out-degree centrality of the citing paper, and others, which demonstrate statistical relevance in classifying citation links. Zhu et al. (2015) analyze the Pearson correlation coefficients between various citation features and the academic influence of references. Their findings suggest that citations placed at the beginning of a sentence and the standard variance of citation locations could be more influential. Qayyum and Afzal (2019) emphasize the potential of metadata parameters, including titles, authors' names, keywords, and categories, in identifying important citations. Lyu et al. (2021) decode citing motivations using standard meta-synthesis procedures.

Despite the progress made in citation analysis based on feature engineering, citation intent analysis task remains complex and difficult to measure by feature engineering alone, as it is a semantic-level task that requires a deep understanding of the text. Although many citation feature sets have been developed for specific fields, feature engineering approaches may reduce generality. Furthermore, there is some underutilized information, including citation section and citation worthiness information has also been proven useful in citation analysis (Cohan et al., 2019; Jochim & Schütze, 2012; Zhang et al., 2022; Zhu et al., 2015).

Citation intent analysis based on deep learning

With the remarkable progress of pretrained language models, deep learning based models were getting more and more attention, owing to their powerful representation of the deep semantic information in citation texts. In recent years, citation studies based on pretrained language models such as BERT (Devlin et al., 2018) have been widely carried out. For instance, Roman et al. (2021) propose a text clustering mechanism to automatically annotate citation intent and find BERT significantly outperforms the Glove word embedding

model. But their clustering annotation method regards the citation intent classification as an unsupervised learning task and updates the citation intent scheme according to the data set, which needs to be further verified. For scientific literature, Beltagy and Cohan (2019) train SciBERT, a language model pretrained on a large scaled multi-domain scientific publications corpus. Maheshwari et al. (2021) prove that SciBERT could capture nuances of scientific documents with BiLSTM and Random Forest, and ranked outstanding on 3C citation context classification shared task at NAACL 2021. Lauscher et al. (2021) demonstrate the importance of using precisely-defined variable-length contexts to pretrain SciBERT and RoBERTa language model for multi-sentence multi-intent citation analysis. However, their goal is to assess the importance of gold citation contexts, rather than to enhance citation classification performance. They rely on the manual annotation to determine gold citation contexts, which limits the feasibility and versatility of their model. To address the complex scenario of document-level multi-label citation function classification, Zhang et al. (2021) propose a two-stage fine-tuning pre-trained citation document representation model. In the first stage, unlabeled data is utilized to enhance the model's awareness of citations and their positions. In the second stage, labeled data is used to train the model with a multi-label loss function. Their study reveals that a relatively small dataset size could lead to unstable performance of pre-trained language models.

It is important to note that deep learning based models require comprehensive data to obtain stable performance. However, the semantic information in citation intent is complex and difficult to annotate automatically. To improve the performance of citation intent analysis with limited annotated data, word embedding and pretrained language models are combined with citation features to represent citation context. For instance, Andrade and Gonçalves (2020) combine contextual information in Glove word embedding, statistical information in TF-IDF, and topical information in the LDA model to classify citation intent and citation influence. Prester et al. (2021) develop a deep content-enriched ideational impact classification model to aid literature searches based on the research of Dong and Schäfer (2011). Zhang et al. (2022) combine native citation features with SciBERT to classify citation function by LSTM deep learning algorithm. Inspired by Prester et al. (2021) and Dong and Schäfer (2011), this paper established a new citation representation model composed of pretrained language models and heterogeneous features as the input of multi-task citation classification framework with soft parameter sharing constraint.

Citation intent analysis based on multi-task learning

Considering the limited amount of available annotated citation intent data, in order to make full use of existing annotation data of related tasks and potential related semantic information between different related tasks, multi-task learning method is applied in citation intent classification. As a promising deep learning method, multi-task learning aims to exploit useful information and enhance semantic information on multiple tasks to improve overall performance (Zhang & Yang, 2018). It has been proven effective on many natural language processing tasks (Qi et al., 2022a; Yousif et al., 2019). Task relatedness and task definition are the two elementary factors of multi-task learning (Zhang & Yang, 2018). To reveal the relationship between citation analysis related tasks, such as citation location, importance, worthiness, and sentiment classification tasks, several studies have been carried out. For citation sentiment and citation purpose classification tasks, Yousif et al. (2019) propose a multi-task learning model based on CNN and RNN deep learning models which benefit from joint learning by modeling the citation context with task-specific information

and shared layers. Their study proves that sentiment information in the citation context is conducive to citation intent analysis tasks. To prove citation function and provenance are closely related tasks, Su et al. (2019) build a CNN deep learning model with Glove embedding to classify citation function and citation provenance simultaneously. More recently, in order to explicitly represent the information on auxiliary tasks, Hu et al. (2022) propose a multi-task model based bilateral-branch network for imbalanced citation intent classification by constructing a shared encoder layer and no-shared encoder layer respectively.

For the citation intent classification task, Cohan et al. (2019) propose a multitask model, Structural Scaffolds, which takes citation location and worthiness classification as auxiliary tasks and incorporates structural information of scientific papers for citation intent classification. They prove that the auxiliary tasks are helpful to citation intent classification. Instead of depending on external linguistic resources or manually engineered features, Structural Scaffolds mainly rely on Glove-ELMo word vector representation and self-attention BiLSTM structural scaffold multi-task learning model. Besides, the constraint of the hard parameter sharing mechanism can be overly restrictive on auxiliary tasks, leading to a performance decline when the tasks are not closely related (Ruder, 2017). Following Cohan et al. (2019), Oesterling et al. (2021) represent the citation text with TF-IDF vectors, manually compiled vocabulary set, counts of citations in each section, citation's relative position, and Glove-ELMo embedding to classify citation intent, citation section and citation worthiness. Their researches find that the generated word vector embedding based on Glove-ELMo and LSTM with attention plays a major role in the citation intent classification task. But this method relies on manually generated vocabulary and ignores syntactic features such as part-of-speech and syntactic patterns that can be automatically generated by natural language processing techniques and have been proved to facilitate citation intent analysis.

Methodology

In this study, we propose a multi-task citation intent classification framework (MTCIC), which consists of the citation intent classification task as the main task, citation worthiness classification and citation section classification as the auxiliary tasks. MTCIC supports diverse inputs for each task, depicted in Fig. 1. In the multi-task framework, firstly we obtain the context semantic representation of full text in citation corpus via SciBERT (Beltagy & Cohan, 2019). Secondly, for the citation intent classification main task,

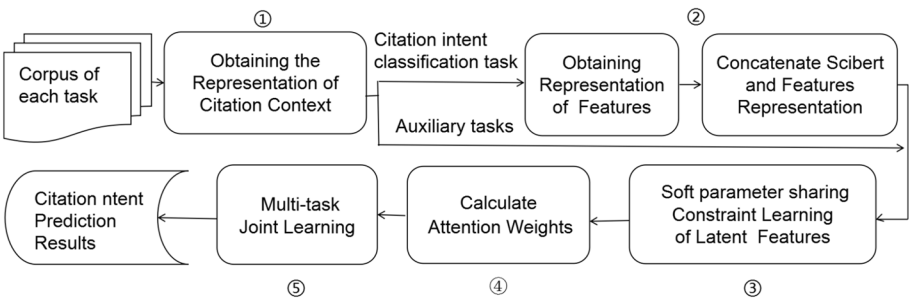


Fig. 1 The main process of multi-task citation intent classification

heterogeneous feature sets are extracted and concatenated with the output of SciBERT to enhance semantic information, thus improving the performance of citation intent classification. The heterogeneous feature sets include syntax, statistical, structural and sentiment features. Thirdly, based on soft parameter sharing constraint, which allows more flexibility and adaptability to task-specific features and requirements by assigning different weights to the shared parameters for each task, representation vectors for each task are learned by BiLSTM respectively as shown in Fig. 2. Fourthly, multi-head attention is introduced to help the proposed model focus on the important information in the input vectors. Then, multiple tasks are jointly learned with Cross Entropy and weighted loss. Finally, the citation intent labels are predicted. The overall goal of this framework is to optimize the learning process of the citation intent classification main task. The detailed description of the above main steps is illustrated as follows.

Joint representation for citation multi-task

In order to obtain complementary information in multi-source heterogeneous citation features, we proposed a joint heterogeneous citation multi-task representation model based on soft parameter sharing constraint. In steps 1 and 2 in Fig. 1, for the citation

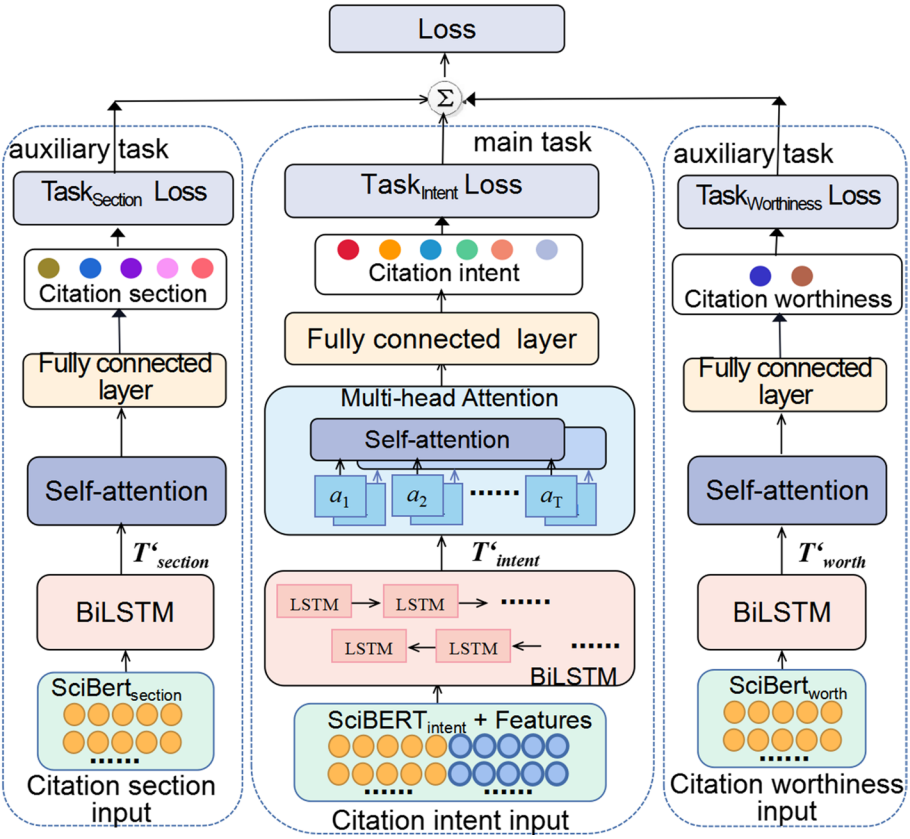


Fig. 2 Multi-task citation intent classification framework

intent classification task, a concatenated feature set is designed as shown in formula (1), including general semantic and scientific domain information from SciBERT, heterogeneous citation information from syntactic, statistical, structural, and sentiment features. For auxiliary tasks, the representation of citation context is generated by a pretrained language model, namely SciBERT, as shown in formulas (2) and (3).

$$\text{Input}_{intent} = \left(\text{SciBERT}(C), feature_{ij}(C)\right) \tag{1}$$

$$\text{Input}_{worth} = \text{SciBERT}(C) \tag{2}$$

$$\text{Input}_{section} = \text{SciBERT}(C) \tag{3}$$

In formula (1), the general semantic and scientific citation information in citation context C is represented via SciBERT and heterogeneous features $feature_{ij}$. Let i be the context number and j be the word number in the context, $feature_{ij}$ is composed of the following features as shown in formula (4). In addition to SciBERT, we include heterogeneous features for predicting the citation intent, described in Eq. (4).

$$feature_{ij} = \text{cat}\left(\left[\text{Onehot}_{ij}\left(pos_{ij}, pos\_list\right), pattern_{ij}, senti_{ij}, tfidf_{ij}\right]\right) \tag{4}$$

In Eq. (4), the cat function is to concatenate the given sequence of features to a new vector. The features in Eq. (4) are as follows:

- The idea of using part of speech tag as a feature is inspired by Teufel et al. (2006). The $\text{Onehot}_{ij}(pos_{ij}, pos\_list)$ is the part of speech expressed in one hot vector.
- The $pattern_{ij}$ is syntactic structures that the citation context contains (Dong & Schäfer, 2011; Jochim & Schütze, 2012; Prester et al., 2021), including the following six syntactic structures: (i) citation + verb [past/present/third person/past participle]; (ii) verb [past/gerund/third person] + verb [gerund/past participle]; (iii) verb [all forms] + (adverb [comparative/superlative]) + verb [past participle]; (iv) modal + (adverb [comparative/superlative]) + verb + (adverb [comparative/superlative]) + past participle; (v) (adverb [comparative/superlative]) + Personal Pronoun + (adverb [comparative/superlative]) + verb [all forms]; (vi) gerund + (proper noun + juxtaposition conjunction + proper noun).
- The sentiment feature $senti_{ij}$ is the weighted embedding representation of the domain sentiment word vector multiplied by TF-IDF vector. The calculation of $senti_{ij}$ is inherited from Prester et al. (2021) and Qi et al. (2022b).
- The idea of $tfidf_{ij}$ as a feature is from Andrade and Gonçalves (2020) and Oesterling et al. (2021). And $tfidf_{ij}$ of each word is calculated as shown in (5), where d is an instance in the citation corpus, $f_{j,d}$ is the frequency of the word j in instance d, N is the number of all instances in the citation corpus, and n_j is the number of instances with the word j.

$$W_{tf-idf} = \log\left(1 + f_{j,d}\right) \times \log\left(\frac{N}{1 + n_j}\right) + 1 \tag{5}$$

Deep learning and multi-head attention

In step 3 in Fig. 1, the deep learning and attention module of each task includes BiLSTM and multi-head attention layers as shown in Fig. 2. In the BiLSTM layer, a preliminary citation context vector $T\prime$ is extracted from SciBERT and the heterogeneous citation feature set. Then, to focus on the most pertinent information for each task, the attention weight matrix of $T\prime$ is calculated by the multi-head or single head self-attention mechanism in the model.

In step 4 in Fig. 1, the self-attention layer, input $T\prime$ is mapped to the query vector Q, the key information K, and the word embedding vector V for each word. The attention is calculated as shown in formula (6), where d_k is the dimension of $T\prime$. The dot product of Q and K generates an attention weight map. And then the attention-weighted features are generated by the dot product of the attention map and V. In the calculation of dot product attention as shown in formula (6), $\frac{1}{\sqrt{d_k}}$ is the scale factor to adjust the dot product and smooth the Softmax function (Vaswani et al., 2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

In the framework of this paper, two auxiliary tasks adopt the self-attention mechanism as shown in formula (6). Considering that the input of the citation intent classification task consists of SciBERT and heterogeneous features, the main task adopts multi-head attention. In addition to improving training efficiency, the multi-head attention mechanism further optimizes the overall training performance by dividing the input vector into multiple sub-spaces and sharing training parameters in these multiple sub-spaces during attention weights computation. The process of multi-head attention calculation in the main task is shown in formula (7), where the main task employs two heads, and the attention on the i-th sub-space is shown in formula (8) (Vaswani et al., 2017), in which W_i^Q, W_i^K and W_i^V is the weight parameter matrices of Query, Key and Value in the i-th subspace, and W^O is the parameter to be learned. We then get a vector representing the whole input sequence of each task after self-attention layer.

$$mht = \text{Compute\_Attention}\left(\text{Concat}\left(head_1, head_2\right)W^O\right) \tag{7}$$

$$head_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{8}$$

Multi-task learning and prediction

In steps 5 and 6 in Fig. 1, the multi-task learning layer, multiple tasks learn features and share parameters from different perspectives, thereby enhancing the generalization ability of the model and the performance of the main task. In the MTCIC framework, multiple tasks learn features and share parameters with soft parameter sharing constraints, as shown in Fig. 2. That is, each task has its own parameters, and finally expresses the similarity by constraining the differences between the parameters of different tasks. For binary classification tasks such as citation worthiness classification, the Sigmoid

function is used as the activation function as shown in formula (9), where (x_0,x_1) is a two-dimensional vector that is mapped by the vector representation of the whole input sequence of the classification task in the deep learning and attention module.

$$\text{Sigmoid}\left(x_0, x_1\right) = \left[\frac{1}{1+e^{-x_0}}, \frac{1}{1+e^{-x_1}}\right] \tag{9}$$

For multi-classification tasks such as citation intent or section classification, the Softmax function is used as the activation function, as shown in formula (10), in which K is the number of intent categories, and x_k is the vector mapped by the vector representation of the whole input sequence of the classification task in the deep learning and attention module.

$$\text{Softmax}\left(x_1 \ldots x_k\right) = \left[\frac{e^{x_1}}{\sum_{j=1}^{K} e^{x_1}} \ldots \frac{e^{x_k}}{\sum_{j=0}^{K} e^{x_k}}\right] \tag{10}$$

Then, learning of each task is constrained by cross-entropy loss function, with the addition of L2 regularization to prevent overfitting, as shown in formula (11), in which p_i is the probability that the instance belongs to the category i, $\bar{p}_i$ is the probability that the instance does not belong to the category i, and w is the parameter of the weight matrix in the fully connected layer. The jointed loss function is shown in formula (12), where $L_j(wtaskj)$ is the loss function of task j, and λ_i is the weight adjustment factor of the multiple tasks.

$$L(w) = -\sum_{i=1}^{n}(\bar{p}_i \log p_i + (1-\bar{p}_i)\log(1-p_i)) + \alpha(||w_1||_2^2 + \ldots| + |w_i||_2^2) \tag{11}$$

$$Loss = \sum_{w \in Task_0} L_0(w_{task0}) + \lambda_1 \sum_{w \in Task_1} L_1(w_{task1}) + \ldots + \lambda_m \sum_{w \in Task_m} L_m(w_{taskm}) \tag{12}$$

Experimentation and results

Dataset

We employed the publicly available citation dataset ACL-ARC (Jurgens et al., 2018) and SciCite (Cohan et al., 2019) as the benchmark dataset to compare our model with previous works. In the ACL-ARC dataset, three tasks (namely, citation intent, citation worthiness and citation section classification task) are manually annotated by domain experts in the NLP field. The annotated data for each task in this corpus are different. The citation intent subset includes a total of 1941 instances, which are divided into six categories according to the ACT annotation system (Pride et al., 2019). There are 50,000 instances in the citation worthiness subset, among which only 14% belong to "True" category, which leads to dataset imbalance. The citation section subset has 47,757 instances, of which 45% belong to the "Introduction" section category among five categories. The SciCite dataset contains 11,020 instances and is annotated with citation intent labels, including "Background", "Method" and "Result comparison", and also with isKeyCitation labels, including "True" and "False". Table 1 shows the data distribution of ACL-ARC and SciCite in detail.

Table 1 Distribution of ACL-ARC (Jurgens et al., 2018) and SciCite dataset (Cohan et al., 2019)

Dataset	Task	Categories	Distribution (%)	#instances
ACL-ARC	Citation intent	Background	51	1021
		Compare/contrast	18	344
		Extends	4	73
		Future work	4	68
		Motivation	5	98
		Uses	19	465
	Citation worthiness	True	14	6999
		False	86	43,001
	Citation Section	Introduction	45	21,498
		Related work	20	9489
		Method	12	5872
		Experiments	19	8860
		Conclusion	4	2038
SciCite	Citation intent	Background	58	6376
		Method	29	3153
		Result comparison	13	1491
	isKeyCitation	True	41	4531
		False	59	6489

Experimental setup

The datasets were split into three sets with 85% of the data used for training and the remaining 15% divided equally into validation set and test set according to Cohan et al. (2019). In the experiments, we used the citation context in the "text" field of ACL-ARC and the "string" field of Scicite. For these citation contexts, we retained the English core vocabulary while removing special symbols. We optimized cross entropy loss using Adam, holding SciBERT weights frozen and applying a dropout of 0.1. We trained with early stopping on the training set (patience of 5) using a learning rate of 0.001, and batch size of 32 for ACL-ARC, batch size of 64 for SciCite. The loss weight parameter λ_i for multiple tasks was tuned by grid search for best performance on the validation subset for multiple tasks respectively. When the weight of the main task was fixed as 1, the grid search of auxiliary tasks' weights was performed using the parameter ranges from 0.01 to 1, and the step size was 0.01. For example, the following hyperparameters were used for the ACL-ARC multitask learning: intent $\lambda_1 = 1$, worthiness $\lambda_2 = 0.1$, section $\lambda_3 = 0.08$. On the SciCite multitask learning, the weight λ_1 of the intent classification task was set to 1, and the weight λ_2 of the isKeyCitation task was 0.6. To avoid the contingency of the result, the reported overall results were average of 10 runs with random seeds. The GPU in our experimental system was NVIDIA TITAN Xp 12G, with CUDA 10.0.130.

To evaluate our proposed model, the multi-task citation intent classification framework (MTCIC), we compare our model with research results on ACL-ARC and SciCite datasets in recent years as follows:

- Jurgens et al. (2018). This baseline introduces a machine learning classifier with pattern-based features, topic-modeling features, citation graph features, section titles and relative section position features to train the citation intent classifier on ACL-ARC dataset and is published in *Transactions of the Association for Computational Linguistics* in 2018.
- Cohan et al. (2019) BiLSTM-Attn with ELMo. This baseline uses Glove word embedding vectors concatenated with ELMo as the input of the BiLSTM network with attention mechanism, optimizing the deep learning network with the loss function of the citation intent classification main task.
- Cohan et al. (2019) Structural-Scaffold. This baseline adopts the Structural-Scaffold multi-task structure based on hard parameter sharing constraints, using Glove word embedding vectors concatenated with ELMo as input. It achieves a new state-of-the-art on ACL-ARC dataset by optimizing the network with multitask learning in NAACL2019.
- Beltagy and Cohan. (2019). This baseline releases fine-tuned SciBERT pre-trained language model and demonstrates statistically significant improvements by using SciBERT as the input of a fully connected layer on the citation intent classification task in EMNLP2019.
- MTCIC with Glove. To compare MTCIC against Cohan et al. (2019) in a fairer way, this model replaces SciBERT in MTCIC with Glove as the word embedding.
- Multi-Task Citation Intent Classification framework (MTCIC). The citation intent, worthiness and section multi-task classification model proposed in this paper, with the composition of SciBERT and heterogeneous features as input.

Results and discussion

In order to compare the overall performance of citation intent classification, the main experimental results of the model proposed in this paper against several benchmarks are shown in Table 2. Firstly, we observed that the MTCIC model proposed in this paper achieves noticeable improvements over the state-of-the-art approaches on the citation

Table 2 Citation intent classification results compared with baselines

Model	ACl-ARC Macro F1%	Scicite Macro F1%
Jurgens (2018) random forest	54.60	79.6
Cohan (2019) BiLSTM-Attn	51.80	77.2
Cohan (2019) BiLSTM-Attn-ELMo	54.30	82.6
Cohan (2019) structural-scaffold	63.1	79.1
This work MTCIC -glove	64.03	79.54
Cohan (2019) BiLSTM-Attn-ELMo-Scaffolds	67.90	84.0
Beltagy (2019) SciBERT Frozen	60.74	85.42
Beltagy (2019) SciBERT Finetune	70.98	85.49
This work MTCIC SciBERT Frozen	**75.78**	**85.54**

Bold highlight the highest number in different index

intent classification task. Overall, MTCIC significantly improved classification performance on key indicators, with macro F1 21.18% higher on ACL-ARC and 5.94% higher on Scicite than Jurgens et al. (2018), macro F1 12.68% higher on ACL-ARC and 6.44% higher on Scicite than Structural-Scaffold (Cohan et al., 2019), macro F1 15.04% higher on ACL-ARC and 0.12% higher on Scicite than Beltagy and Cohan's SciBERT Frozen (2019). It should be noted that the model's enhancement on Scicite is less apparent compared to ACI-ARC. The possible reason for it could be the overfitting caused by training a deep learning model on small datasets. The performance of the deep learning model heavily relies on the quantity of the training data. When the training dataset is small, the model may inadvertently learn noise in the data, resulting in limited generalization to new data. Another possible reason is that the Scicite dataset provides annotations only on the sentences where citations appear, while existing research suggests that citation function classification should be conducted on the entire citation context rather than individual citation sentences. (Jiang & Chen, 2023). On the ACI-ARC dataset, the annotated data for both subtasks amounts to 97,757 instances. The main task consists of 1941 instances, while the auxiliary tasks encompass a significant amount of additional data that is not covered by the main task. The quantity of data in the auxiliary tasks far exceeds that of the main task, with the data volume being more than 50 times larger. This abundance of data includes valuable information not present in the main task alone, thus the inclusion of auxiliary tasks significantly improves the performance of citation intention classification. In contrast, the main task on the Scicite included 11,020 instances, and the quantity of data in the auxiliary task is the same as main task. Besides, the context of the auxiliary task is similar to that of the main task, providing limited additional valuable information. Consequently, the similarity between the main and auxiliary tasks can lead to overfitting and diminish the model's performance. Similarly, it can be observed that in the experiments conducted by Cohan et al. (2019), the Structural-Scaffold multi-task model exhibits a significant improvement in citation intention classification performance on ACl-ARC, with an increase of 8.8%. However, on the Scicite dataset, the introduction of the auxiliary task results in a decrease of 3.5% in the F1 score.

In order to compare MTCIC against Cohan et al. (2019) in a fair way, we replaced SciBERT in MTCIC with Glove word embedding, and found the macro F1 of MTCIC with Glove still 0.93% higher on ACL-ARC and 0.44% higher on Scicite than Structural-Scaffold (Cohan et al., 2019). This result indicates that compared with the Structural-Scaffold (Cohan et al., 2019) which is based on word vector representation and hard parameter sharing constraint multi-task framework, the joint heterogeneous representation and multi-task framework with soft parameter sharing constraint proposed in this paper provides major contributions to citation intent classification. In addition, the MTCIC with SciBERT frozen has the obvious advantage of macro F1 compared with SciBERT fine-tuned model (Beltagy & Cohan, 2019) ($\Delta = 4.8\%$ on ACL-ARC) proving that the improvement in citation intent classification performance mainly comes from the heterogeneous features and soft parameter sharing constraint multi-task framework, instead of SciBERT pretrained language model.

In order to investigate the effect of the auxiliary tasks on the main task, the results of task ablation experiments on two datasets are shown in Tables 3 and 4. In the auxiliary task ablation experiments on ACL-ARC, removing the citation section auxiliary task resulted in the F1 score decline from 75.78% to 73.27% ($\Delta = 2.51\%$). Removing the citation worthiness auxiliary task caused the F1 score's decline to 73.07% ($\Delta = 2.71\%$). When both auxiliary tasks were removed, F1 decreased to 72.75% ($\Delta = 3.03\%$). Similarly, on Scicite, removing the isKeyCitation auxiliary task caused the F1 score's

Table 3 Task ablation experiment On ACL-ARC

Tasks in MTCIC	(Macro Avg %)		
	Precision	Recall	F1
Intent + section + worthiness	**82.07**	**72.80**	**75.78**
Intent + worthiness	76.38	72.58	73.27
Intent + section	77.03	71.77	73.07
Intent	77.05	71.71	72.75

Bold highlight the highest number in different index

Table 4 Task ablation experiment On Scicite

Tasks in MTCIC	(Macro Avg %)		
	Precision	Recall	F1
Intent + isKeyCitation	**86.48**	**85.20**	**85.79**
Intent	81.99	83.67	82.67

Bold highlight the highest number in different index

Table 5 Function module ablation experiment

Modules in MTCIC	ACL-ARC (Macro Avg %)			Scicite (Macro Avg %)		
	Precision	Recall	F1	Precision	Recall	F1
SciBERT + Attention + BiLSTM + Features	77.05	**71.71**	**72.75**	**81.99**	**83.67**	**82.67**
SciBERT + Attention + BiLSTM	76.77	68.78	71.12	81.70	82.54	81.90
SciBERT + Attention	**83.92**	63.19	68.82	73.64	76.28	71.74
SciBERT	82.66	62.68	68.26	72.68	74.11	69.25

This is citation intent single task experiment. Bold highlight the highest number in different index. SciBERT is frozen in MTCIC

decline from 85.79% to 82.67% ($\Delta = 3.12\%$). This indicates that both auxiliary tasks respectively provide complementary information that is useful for citation intent prediction, and combining them with the main task provides effective complementary information for citation intent prediction.

In order to investigate the effect of each function module in the proposed framework, the results of function module ablation experiments on two datasets are shown in Table 5. In the function module ablation experiments, removing the heterogeneous features of MTCIC decreased the F1 from 72.75% to 71.12% ($\Delta = 1.63\%$) on ACL-ARC and F1 from 82.67% to 81.90% ($\Delta = 0.77\%$) on Scicite, suggesting that the heterogeneous features provide useful information for citation intent prediction. Removing the BiLSTM layer of MTCIC led to a decline in F1 to 68.82% ($\Delta = 2.30\%$) on ACL-ARC and a decline in F1 to 71.74% ($\Delta = 10.16\%$) on Scicite. When we used the attention mechanism alone, the precision of ACL-ARC increased, which may be due to the precision increase in some categories. But the recall and F1 decreased, indicating a decline in overall classification performance. Removing the attention layer of MTCIC led to a decline in F1 to 68.26% ($\Delta = 0.56\%$) on ACL-ARC and a decline in F1 to 69.25% ($\Delta = 2.49\%$) on Scicite. It suggests that the BiLSTM layer and attention mechanism help to extract implicit information on multiple feature spaces for citation intent prediction.

Detailed per category results

Because the distribution of six citation intent labels in ACL-ARC is obviously unbalanced as shown in Table 1, in the experiment result, the Micro-F1, precision or recall is mainly dominated by the categories with more instances. To investigate the performance of the classifier on each category, we picked the models and reported the corresponding performance across all labels in ACL-ARC as shown in Table 6, where the top two Micro-F1scores on each category are shown in bold. It can be observed that in contrast to baseline models, the proposed model achieved higher Marco-F1 scores in most of the categories, including the "Background" category and minority categories such as "Extension", "Future", "Motivation" and "Use". Most of the baseline models, such as Structural-Scaffold, achieve higher F1 in the "Background" category, but obviously lower Marco-F1 scores in minority categories.

In the ablation models, when the heterogeneous features were removed from MTCIC, the Marco-F1 scores of the "Extension", "Future" and "Use" categories were significantly reduced, proving heterogeneous features help citation intent recognition. When the citation worthiness classification auxiliary task was removed from MTCIC, the Marco-F1 scores of minority categories were significantly reduced. When the citation section classification auxiliary task was removed from MTCIC, the Marco-F1 scores of the "Future", "Motivation" and "Use" categories were significantly reduced. When the multi-head mechanism was removed from MTCIC, the Marco-F1 scores of the "Extension", "Future" and "Use" categories were significantly reduced. It indicates that the multi-headed attention mechanism improves the recognition of minor citation intent categories.

To investigate the performance of the multi-task model in dealing with unbalanced data sets, the confusion matrix of MTCIC, MTCIC-without-features and MTCIC-without-auxiliary-tasks are picked and shown in Fig. 3. As can be seen from Fig. 3, generally, the misclassification to major categories in Cohan et al.'s (2019) model was improved significantly on MTCIC. However, there were still a few "Background" and "Compare" instances misclassified to each other. Several instances of "Background", "Compare" and "Future" were misclassified to "Use", and the error instances of "Motivation" were mainly misclassified to "Compare" or "Background". When the heterogeneous features were removed from MTCIC, "Background" was more easily confused with "Extension", "Future" or "Motivation". And there was more confusion between "Compare" and "Background". Instances of "Future" were more likely misclassified to "Background". Instances of "Use" had more misclassification to "Background", "Compare" and "Future". When the auxiliary tasks were removed from MTCIC, the true positive rate of "Background", "Compare" and "Use" was lower, and "Use" was misclassified to "Extension" and "Future" more. However, the single task model of MTCIC performed better on "Extension", possibly because the information from two auxiliary tasks has slight interference in the judgment on the "Extension" category.

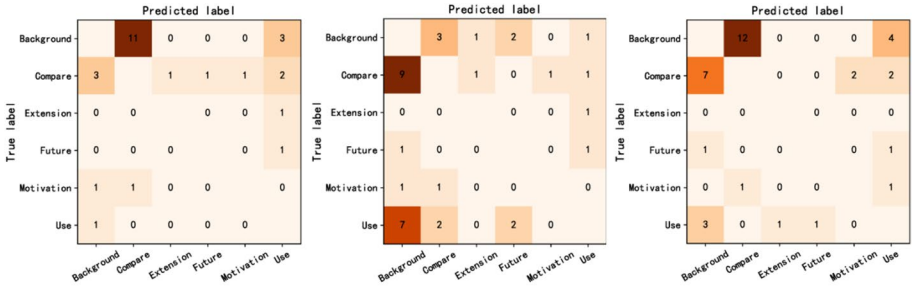
Case study

To gain more insight into how the multi-task mechanism helps the MTCIC improve citation intent classification performance, we examined the attention weights assigned to the inputted instances. We conducted this case study on the following instance from the ACL-ARC dataset. The label of this instance is ''Background''.

Table 6 Detailed per category classification results

Model	Background (71)			Compare (25)			Extension (5)			Future (5)			Motivation (7)			Use (26)		
	P %	R %	F1%	P %	R %	F1%	P %	R %	F1%	P %	R %	F1%	P %	R %	F1%	P %	R %	F1%
Jurgens R-Forest	75.6	87.3	81.1	70.6	48.0	57.1	66.7	40.0	50.0	50.0	20.0	28.6	75.0	42.9	54.6	51.6	61.5	56.1
CohanBiLSTM-Att	76.5	87.3	81.6	59.1	52.0	55.3	66.7	40.0	50.0	33.3	40.0	36.4	50.0	28.6	36.4	69.6	61.5	65.3
CohanStru-Scaffold	75.9	93.0	83.5	80.0	64.0	**71.1**	75.0	60.0	66.7	75.0	60.0	66.7	100	28.6	44.4	81.8	69.2	75.0
MTCIC	91.9	80.3	**85.7**	58.6	68.0	63.0	80.0	80.0	**80.0**	80.0	80.0	**80.0**	83.3	71.4	**76.9**	78.1	96.2	**86.2**
Without features	78.0	90.1	83.7	68.4	52.0	59.1	66.7	80.0	72.7	42.9	60.0	50.0	83.3	71.4	**76.9**	78.9	57.7	66.7
Without section	87.0	84.5	**85.7**	63.6	56.0	59.6	80.0	80.0	**80.0**	60.0	60.0	60.0	62.5	71.4	66.7	66.7	76.9	71.4
Without worthi	84.3	83.1	83.7	55.6	60.0	57.7	60.0	60.0	60.0	100	60.0	**75.0**	80.0	57.1	66.7	72.4	80.8	76.4
Without auxi	83.3	77.5	80.3	51.9	56.0	53.8	83.3	100	**90.9**	75.0	60.0	66.7	71.4	71.4	**71.4**	72.4	80.8	76.4
Without m-head	84.4	91.5	**87.8**	77.8	56.0	**65.1**	100.0	40.0	57.1	100	60.0	**75.0**	60.0	85.7	70.6	75.9	84.6	**80.0**

Bold highlight the highest two number in different index



(a) Confusion matrix of MTCIC (b) MTCIC without features (c) MTCIC without auxiliary tasks

Fig. 3 Confusion matrix of citation intent classification results

We will examine the worst-case complexity of interpretation as well as generation to shed some light on the hypothesis that vague descriptions are more difficult to process than others because they involve a comparison between objects (Beun and Cremers 1998, Krahmer and Theune 2002).

Figure 4 shows the citation along with the horizontal line and the heat map of attention weights for this input instance resulting from MTCIC with multi-task versus MTCIC with single-task. It can be observed from Fig. 4 that in MTCIC with the multi-task model, more weight is placed on the words surrounding "generation to shed" and "comparison between objects," aiding in obtaining a more comprehensive understanding of the sentence's semantic information. On the other hand, the MTCIC with the single-task model attends most to the words "examine the worst-case" and consequently incorrectly predicts a "Compare" label. Note that the only difference between these two models is the auxiliary tasks. It demonstrates the auxiliary tasks provide relevant signals for citation intent classification main task.

Conclusion and future work

Citation text is the evaluation and interpretation made by the author when citing a document, bearing the intent and emotional bias of the author. In this paper, we proposed a Multi-Task Citation Intent Classification framework that uses the soft parameter sharing mechanism to constrain the relationship between multiple tasks learning models. To improve the performance of the citation intent classification task, we also proposed a

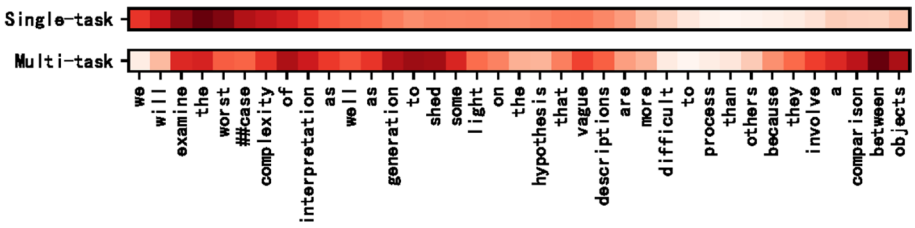


Fig. 4 Visualization of attention weights corresponding to our best MTCIC model compared with the MTCIC model without auxiliary tasks

joint citation representation model for the MTCIC framework, including general semantic and scientific domain information from the pretrained language model and heterogeneous features.

Our experimental results suggest that the proposed MTCIC framework and joint citation representation model outperforms contrast methods, and the auxiliary tasks and joint citation representation model do contribute to the citation intent classification. In the detailed per category results and case study, we further verified that auxiliary tasks and heterogeneous features help to improve the recognition of minor citation intent categories. The proposed MTCIC would facilitate the evaluation the quality of the literature and its academic impact. It also contributes to deep scientific research behavior evaluation, bibliometric evaluation, citation information retrieval, recommendation and prediction. Furthermore, this study helps to extract and manage knowledge entities, including but not limited to data sets, knowledge elements, methods, tools and theories. Our findings provide insights into the design of effective citation intention classification models.

However, it is important to acknowledge that the proposed model does have certain limitations. The performance of the model could be influenced by the quality and quantity of annotated data available for each task. Additionally, our study is constrained by the quality and scope of the features extracted from individual datasets for each task. Moreover, it is worth noting that the proposed model may not be directly applicable to other languages, as it has been specifically trained and evaluated on English language datasets.

These limitations present opportunities for future research to investigate and address. In our future work, we aim to enhance our model by further exploring the feature extraction methods and examining the relevance of additional auxiliary tasks. Furthermore, we plan to delve into the mechanism of the parameter sharing constraints in the multi-task learning framework, with the goal of improving the performance of the multi-task models specifically in the domain of citation intent analysis. Additionally, we will explore the performance and potential improvements of our model in multilingual environments.

Appendix

See Tables 7 and 8.

Table 7 Experiment results of using feature set as the input of the single task of citation intention classification

Features	ACL-ARC (Macro %)			Scicite (Macro %)		
	Precision	Recall	F1	Precision	Recall	F1
Pos	49.98	43.31	44.62	**72.36**	62.39	65.10
Pos + PosPattern	53.39	46.69	47.55	71.42	62.72	65.23
Pos + PosPattern + Senti	**61.40**	48.03	50.71	71.87	62.67	65.49
Pos + PosPattern + Senti + Tfidf	59.12	**49.77**	**51.62**	66.33	**66.51**	**66.12**
Pos + PosPattern + Senti + Tfidf + SectionName	55.51	46.94	48.65	65.93	63.77	64.74
Pos + PosPattern + Senti + Tfidf + SectionName + Offset	52.53	47.58	48.12	19.61	33.33	24.67

Bold highlight the highest number in different index

Table 8 Experimental results of different auxiliary tasks on SciCite dataset

Tasks	Task description	Categories	Distribution (%)	#Instances	MTCIC SciBERT Frozen Scicite (Macro F1%)
Single task	Citation intent total# 11020	Background	58	6376	82.67
		Method	29	3153	
		Result comparison	13	1491	
One auxiliary task	isKeyCitation Total# 11020	True	41	4531	85.54
		False	59	6489	
Two auxiliary tasks	Citation Worthiness Total# 73484	True	14	10,532	**86.28**
		False	86	62,952	
	Citation section total# 91412	Introduction	39	35,312	
		Related work	8	7454	
		Method	16	14,774	
		Experiments	16	14,991	
		Conclusion	21	18,881	

Bold highlight the highest number in F1

Acknowledgements This work is partially supported by grant from the Applied Basic Research Project of Liaoning Province (No. 2022JH2/101300270), the Scientific Research Innovation Team Project of Dalian University of Foreign Languages (No. 2016CXTD06)

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. Preprint at http://arXiv.org/arXiv:1903.10676

Cohan, A., Ammar, W., Van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. Preprint at http://arXiv.org/arXiv:1904.01608

de Andrade, C. M. V., & Gonçalves, M. A. (2020). Combining representations for effective citation classification. *In Proceedings of the 8th International Workshop on Mining Scientific Publications*: 54–58.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at http://arXiv.org/arXiv:1810.04805

Dong, C., Schäfer, U.(2011). Ensemble-style self-training on citation classification, *Proceedings of the 5th International Joint Conference on Natural Language Processing*. 623–631.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science, 178*(4060), 471–479.

Hassan, N. R., & Serenko, A. (2019). Patterns of citations for the growth of knowledge: A Foucauldian perspective. *Journal of Documentation., 75*(3), 593–611.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*(3), 1645–1662.

Hu, T., Li, J., Fukumoto, F., & Zhou, R. (2022). A multi-task based Bilateral-Branch Network for imbalanced citation intent classification. *In 2022 16th International Conference on Ubiquitous Information Management and Communication*. 1–8.

Jiang, X., & Chen, J. (2023). Contextualised segment-wise citation function classification. Scientometrics, 1–42.

Jochim, C., & Schütze, H. (2012). Towards a generic and flexible citation classifier based on a faceted classification scheme. *In Proceedings of COLING*. 1343-1358

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics, 6*, 391–406.

Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Jurgens, D., Cohan, A., & Lo, K. (2021). MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. Preprint at http://arXiv.org/arXiv-2107

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021). The classification of citing motivations: A meta-synthesis. *Scientometrics, 126*(4), 3243–3264.

Maheshwari, H., Singh, B., & Varma, V. (2021). Scibert sentence representation for citation context classification. *In Proceedings of the Second Workshop on Scholarly Document Processing*. 130–133.

Oesterling, A., Ghosal, A., Yu, H., Xin, R., Baig, Y., Semenova, L., & Rudin, C. (2021). Multitask learning for citation purpose classification. Preprint at http://arXiv.org/arXiv:2106.13275

Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing & Management, 26*(1), 171–186.

Prester, J., Wagner, G., Schryen, G., & Hassan, N. R. (2021). Classifying the ideational impact of information systems review articles: A content-enriched deep learning approach. *Decision Support Systems, 140*, 113432.

Pride, D., Knoth, P., & Harag, J. (2019). ACT: an annotation platform for citation typing at scale. *In ACM/IEEE Joint Conference on Digital Libraries*. 329–330.

Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics, 118*(1), 21–43.

Qi, R. H., Wei, J., Shao Z., Guo X., Chen H. (2022b). Domain Sentiment Lexicon Representation Learning Based on Multi-source Knowledge Fusion. *In Proceedings of the 21st Chinese National Conference on Computational Linguistics*, 684–693. https://aclanthology.org/2022.ccl-1.61/

Qi, R. H., Yang, M. X., Jian, Y., Li, Z. G., & Chen, H. (2022a). A Local context focus learning model for joint multi-task using syntactic dependency relative distance. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03684-0

Roman, M., Shahid, A., Khan, S., Koubaa, A., & Yu, L. (2021). Citation intent classification using word embedding. *IEEE Access, 9*, 9982–9995.

Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. Preprint at http://arXiv.org/arXiv1706.05098

Su, X., Prasad, A., Kan, M. Y., & Sugiyama, K. (2019). Neural multi-task learning for citation function and provenance. *In ACM/IEEE Joint Conference on Digital Libraries*. 394–395.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics, 28*(4), 409–445.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *In Proceedings of the 2006 conference on empirical methods in natural language processing*. 103–110.

Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S. U., & Haddawy, P. (2019). Automatic classification of algorithm citation functions in scientific literature. *IEEE Transactions on Knowledge and Data Engineering, 32*(10), 1881–1896.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *In Workshops at the twenty-ninth AAAI conference on artificial intelligence* (15): 13

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, H., Martin, E., & Mahidadia, A. (2013). Using heterogeneous features for scientific citation classification. *In Proceedings of the 13th conference of the Pacific Association for Computational Linguistics.*

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing, 335*, 195–205.

Zhang, Y., Wang, Y., Sheng, Q. Z., Mahmood, A., Emma Zhang, W., & Zhao, R. (2021). TDM-CFC: Towards Document-Level Multi-label Citation Function Classification. *In International Conference on Web Information Systems Engineering* (pp. 363–376).

Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review, 5*(1), 30–43.

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*. https://doi.org/10.1007/s11192-021-04242-0

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology, 66*(2), 408–427.