Noname manuscript No.
(will be inserted by the editor)

Evaluating BERT-based Scientific Relation Classifiers for Scholarly Knowledge Graph Construction on Digital Library Collections

Ming Jiang · Jennifer D'Souza · Sören Auer · J. Stephen Downie

arXiv:2305.02291v1 [cs.DL] 3 May 2023

Received: date / Accepted: date

Abstract The rapid growth of research publications has placed great demands on digital libraries (DL) for advanced information management technologies. To cater to these demands, techniques relying on knowledge-graph structures are being advocated. In such graph-based pipelines, inferring semantic relations between related scientific concepts is a crucial step. Recently, BERT-based pre-trained models have been popularly explored for automatic relation classification. Despite significant progress, most of them were evaluated in different scenarios, which limits their comparability. Furthermore, existing methods are primarily evaluated on clean texts, which ignores the digitization context of early scholarly publications in terms of machine scanning and optical character recognition (OCR). In such cases the texts may contain OCR noise, in turn creating uncertainty about existing classifiers' performances. To address these limitations, we started by creating OCR-noisy texts based on three clean corpora. Given these parallel corpora, we conducted a thorough empirical evaluation of eight BERT-based classification models by focusing on three factors: (1) BERT variants; (2) classification strategies; and, (3) OCR noise impacts. Experiments on clean data show that the domain-specific pre-trained BERT is the best variant to identify scientific relations. The strategy of predicting a single relation each time outperforms the one simultaneously identifying multiple relations in general. The optimal classifier's performance can decline by around 10% to 20% in F-score on the noisy corpora. Insights discussed in this study can help DL stakeholders select techniques for building optimal knowledge-graph-based systems.

Keywords Digital library · Information extraction · Scholarly text mining · Semantic relation classification · Knowledge graphs · Neural machine learning.

Grants: This material is based upon work supported by the National Science Foundation under Grant No. OAC 1939929 and by the European Research Council for the project ScienceGRAPH (Grant agreement ID: 819536).
Conflicts of interest: Not applicable
Data/Code availability: The clean corpora and the source codes of selected classification models are publicly available. The corresponding noisy corpora and related codes will be released after the paper is published.
Ethics approval: Not applicable
Consent to participate: Not applicable
Consent for publication: Not applicable

M. Jiang
University of Illinois at Urbana-Champaign, USA
E-mail: mjiang17@illinois.edu

J. D'Souza
TIB Leibniz Information Centre for Science and Technology
Hannover, Germany

S. Auer
TIB Leibniz Information Centre for Science and Technology
L3S Research Center at Leibniz University of Hannover
Hannover, Germany

J.S. Downie
University of Illinois at Urbana-Champaign, USA

1 Introduction

Digital libraries (DL) play an important role in promoting scholarly knowledge dissemination and exploration, which is a critical part of scholarly communication. While the accessibility of massive scholarly records in DL brings practical benefits to a variety of research communities for scholarship development, this phenomenon presents new challenges to digital librarians and data curators in managing scholarly information from various sources. In particular, existing modes of document-based scholarly communication create challenges for researchers looking to obtain comprehensive, fine-grained and context-sensitive scholarly knowledge for their specific research topics, especially for multidisciplinary research [23]. According to [4,5], current keyword-

based methods for indexing scholarly articles cause related documents to be scattered and emphasize concentration on selected keywords rather than all aspects of knowledge in an article. Given these limitations, in order to optimize scholarly knowledge organization and representation in DL, some initiatives [23,52] advocate for building an interlinked and semantically rich knowledge graph structure using a combination of human curation and machine learning techniques.

In the process of building knowledge graphs based on scholarly publications, one of the key steps requires establishing semantic relationships between identified scientific terms. Given that paraphrasing is a common phenomenon in natural languages, many identified semantic relationship instances essentially indicate the same relation by different expressions, which can lead to information redundancy in the constructed graphs over large corpora. To avoid this issue, the task of classifying scientific relations (i.e., identifying the appropriate relation type for each related concept pair from a set of predefined relations) is indispensable. Recently, some researchers in the natural language processing (NLP) community have defined seven regular scientific relation types based on their empirical annotations on scholarly articles [6,16], which include HYPONYM-OF, PART-OF, USAGE, COMPARE, CONJUNCTION, FEATURE-OF, and RESULT. The annotations are in the form of generalized relation triples: ⟨experiment⟩ COMPARE ⟨another experiment⟩; ⟨method⟩ USAGE ⟨data⟩; ⟨method⟩ USAGE ⟨research task⟩.

In this age of the "deep learning tsunami", prior work has started to take advantage of powerful neural network techniques to build scientific relation classifiers for the improvement of performance [32]. With the recent introduction of transformer models such as BERT (i.e., Bidirectional Encoder Representations from Transformers) [13] models, the opportunity to obtain boosted machine learning systems is further accentuated. While prior works [8,50] have achieved high classification performance, the results reported in these studies are somewhat incomparable because the proposed methods are usually evaluated on different evaluation corpora. This issue leads to a difficulty in obtaining comparable results and conclusive insights about the effectiveness of developed classifiers in real-world practice. In particular, in the context of academic DLs, digital librarians might find it hard to select the optimal toolkit to satisfy their needs for scholarly knowledge organization based on the findings of prior evaluations (e.g., the underlying data could be different in content, scale, and diversity). Moreover, existing studies focus primarily on the development of techniques for scientific relation classification in the context of unscanned clean corpus [8,30,16], suggesting that researchers tend to concentrate on the ability of machines to automatically identify scientific concepts' relationships from recent scholarly publications that are born in the digital format. However, in broader practice, there exists a wide range of DL collections that were originally published in print and later digitized by machine scanning and OCR. These texts inevitably involve unique digitization noise such as OCR errors, which could challenge the advanced learning-based relation classification techniques. This gap raises further uncertainties about the robustness of state-of-the-art classification models to handle text noise caused by digitization when making predictions.

To address the aforementioned limitations, we focused on providing a comprehensive examination of state-of-the-art BERT-based classification models in a comprehensive and comparable environment considering both clean and OCR-noise data scenarios. With the goal of helping stakeholders within DL select the optimal tool for building scholarly knowledge graphs, we propose a two-stage examination pipeline to: (1) identify the optimal model setting for clean texts; and, (2) further investigate the resilience of the identified optimal model setting for noisy texts with OCR errors, which is one of the most universal limitations in DL collections. Regarding the model setting, we particularly focused on the impact of two key factors: (1) classification strategies (i.e., predicting either a single relation or multiple relations at one time); and, (2) BERT model variants with respect to the domain (i.e., generic or scientific texts) and vocabulary case (i.e., cased or uncased) of their pre-training corpus. As to the measurement of the optimal model's resilience to OCR noise, we started with generating aligned noisy-clean instance pairs by randomly adding a controlled ratio of OCR errors into each clean instance based on an existing word-level OCR-error dictionary, which was provided by HathiTrust Research Center [9]. Given the parallel data, we then compared the optimal model's performance differences on noisy versus clean texts.

Considering that real-world DLs may have various data settings and that such diversity would probably influence the model's performance, we assessed the performance of each model on three corpora including: (1) a single-domain corpus with sparse relation annotations on scholarly publication abstracts in the NLP area [16]; (2) a multiple-domain corpus covering more abundant relations annotated on the publication abstracts from various artificial intelligence (AI) conference proceedings [30]; and, (3) a combination of previous two corpora where the distribution of data domains are unbalanced and annotations are provided by two different groups of annotators. The motivation in building this corpus was to simulate real data settings in digital libraries. With a similar concern, we prepared two noisy versions of each clean corpus with an emphasis on the amount of OCR errors, which included a low-noise version (18%) and a high-noise version (49%).

Experiments on three different clean evaluation corpora showed that the uncased BERT model pre-trained on schol-

arly domain-specific texts was the best variant to classify the type of scientific relations. Regarding the classification strategies, in general, the strategy of predicting a single relation each time achieved a higher classification accuracy than the one identifying multiple relation types simultaneously. By further examining the optimal resulting classifier's performances on three corresponding noisy corpora, we found that the classifier's predictability clearly decreased between 10% to 20% in F-score, indicating that the decreasing rate is even higher with the increasing of the amount of OCR errors in the corpus.

In summary, following our examination pipeline, we addressed the following research questions in this study:

RQ1: What is the performance of eight BERT-based classifiers for scientific relation classification on clean data?
RQ2: Which of the seven studied relation types were easy to identify by the selected optimal classifiers?
RQ3: What kinds of prediction errors are frequently made?
RQ4: How do OCR errors impact the overall optimal classifier's performances?
RQ5: Which of the seven studied relation types are more robust to OCR errors?

2 Related Work

2.1 Relations Mined from Scientific Publications

Overall, knowledge is organized in digital libraries based on two main aspects of a digital collection: (1) metadata; and, (2) content [44,21]. The latter aspect can be further divided into free-form content and ontologized content (i.e., axiomatically defined formal content). In this context, the main categories of relations explored in scholarly publications included two groups.

One group includes metadata relations such as authorship, co-authorship, and citations [49,43]. Research in this group has focused mainly on examining the social dimension of scholarly communications, such as research impact assessments [36,40,18,10,12,34,46,56], co-author prediction [43,47,20,51,37] and scholarly community analysis [49]. Popular data resources that have been explored in this group include Microsoft Academic Graph [42] and PubMed [1].

The second group includes content-based semantic relations, either as semantic categories empirically defined on the basis of the given free-form content [28,26] or as semantic categories formally defined by a systematic conceptual analysis of the properties of concepts within a subject area [41,38]. In the framework of automatic systems, free-form content-based relations have been examined in terms of: (1) relation identification (i.e., recognize related scientific term pairs) [16,26]; and (2) relation classification (i.e., determine the relation type of each term pair, where the relation types are typically pre-defined) [50,8,30]. With respect to ontology-defined properties, prior work primarily considers the conceptual hierarchy based on formal conceptual analysis [41,38].

We attempt to classify free-form content-based semantic relations. Given that digital libraries are interested in the creation of linked data [19], our attempted task directly facilitates the creation of scholarly knowledge graphs and offers structured data to support librarians in generating linked data.

2.2 Techniques Developed for Semantic Relation Classification

Both rule-based [2] and learning-based [11,54] methods have been developed for relation classification. Traditionally, learning-based methods typically relied on hand-crafted semantic and/or syntactic features [2,11]. Among various methods, the strategy of applying distant supervision based on a knowledge database [33,39,22] has been widely used and followed for further improvement. The major advantage of this method is its benefits in solving the challenge of the lack of hand-labeled ground truth for model training.

In recent years, deep learning techniques have been popularly studied because they can more effectively learn latent feature representations for distinguishing between relations. An attention-based bidirectional long short-term memory network (BiLSTM) [54] is one of the first top-performing systems that leveraged neural attention mechanisms to capture important information per sentence for relation classification. Another advanced system [31] leverages a dynamic span graph framework based on BiLSTMs to simultaneously extract terms and infer their pairwise relations. Aside from these neural methods considering the word sequence order, transformer-based models such as BERT [13] that use self-attention mechanisms to quantify the semantic association of each word to its context have become the current state-of-the-art in relation classification. In addition to the generic BERT models trained on books and Wikipedia, recently, Beltagy et al. [8] have developed SCIBERT which are BERT models trained on scholarly publications.

With respect to the classification strategy, prior work regularly adopted a single-relation-at-a-time classification (SRC) that identifies the relation type for an entity pair each time [54,31,8]. To improve the classification efficiency, Wang et al. [50] designed a BERT-based classifier that could recognize multiple pairwise relationships at one time, which can be regarded as a multiple-relations-at-a-time classification (MRC). As opposed to prior work that emphasizes classification improvement, we focus on providing a fine-grained analysis of existing resources for selecting the proper tool

to extract and organize scientific information in digital libraries.

2.3 Relation Classification on Noisy Data

In general, prior work concentrating on the classification of entity relations on noisy data usually focused on noisy annotations, either at entity level such as uncorrected entity boundaries [16,17] or at relation level such as wrong relation labels [25,15,53]. Such noise usually comes from two sources: (1) the biases of human annotations caused by differences in personal understanding of the fine-grained text semantics; and (2) the errors of machine labeling based on distant supervision relying on a generic knowledge base, in which the target content's specific contextual information is missing.

To address this noise in labeling, some methods focus on developing an attention-based neural network model based on the distant supervision learning mechanism, in which the model can learn to combine the generic structural information from existing large-scale knowledge pools with the corpus-specific semantic information, and hence, improve the robustness of relation classifiers [17,25]. Otherwise, some studies [15,53] propose to take advantage of a reinforcement learning strategy: to first select high-quality labeled data, and then feed the selected instances into a relation classifier for training.

Although prior work has made remarkable contributions to improve learning models' robustness in the face of noisy labeling, the text data that these studies rely on is still clean. Given that, our consideration of noisy data is different from prior work in that we concentrate on content-based noise in texts with a specific focus on the OCR content of digitized library collections.

2.4 Impact of OCR Errors on Downstream NLP tasks

With the increasing popularity of applying NLP techniques to DL textual resources for macro-level computation research [24,4], concerns about the reliability of NLP techniques for processing digitized library collections have recently been on the rise [27,45,48]. Based on our literature review, one of the major issues that challenges NLP techniques' reliability on OCR'd texts is their potential inclusion of errors resulting from the OCR process [27,45,48].

According to [14], common OCR errors include character exchange, separated words, joined words, and erroneous symbols. Given the ubiquity, uneven distribution, and heterogeneous nature of OCR errors, it is difficult to fully clean such text noise even using state-of-the-art OCR correction techniques. Given that, there exists an uncertainty about the performance of standard NLP techniques applied on texts with OCR errors.

Overall, existing work concentrating on the investigation of the impact of OCR errors on NLP tasks can be divided

Id	Relation Type	SemEval18		SciERC		Combined	
		Total	%	Total	%	Total	%
1	**USAGE**: a scientific entity that is used for/by/on another scientific entity. E.g. *MT system* is applied to *Japanese*	658	42.13%	2,437	52.43%	3,095	49.84%
2	**FEATURE-OF**: An entity is a characteristic or abstract model of another entity. E.g. *computational complexity* of *unification*	392	25.10%	264	5.68%	656	10.56%
3	**CONJUNCTION**: Entities that are related in a lexical conjunction i.e., with 'and' 'or'. E.g. videos from *Google Video* and a *NatGeo documentary*	-	-	582	12.52%	582	9.37%
4	**PART-OF**: scientific entities that are in a part-whole relationship. E.g. describing the processing of *utterances* in a *discourse*	304	19.46%	269	5.79%	573	9.23%
5	**RESULT**: An entity affects or yields a result. E.g. With only 12 *training speakers* for SI recognition , we achieved a 7.5% *word error rate*	92	5.89%	454	9.77%	546	8.79%
6	**HYPONYM-OF**: An entity whose semantic field is included within that of another entity. E.g. *Image matching* is a problem in *Computer Vision*	-	-	409	8.80%	409	6.59%
7	**COMPARE**: An entity is compared to another entity. E.g. *conversation transcripts* have features that differ significantly from *neat texts*	116	7.43%	233	5.01%	349	5.62%
	Overall	1,562	100%	4,648	100%	6,210	100%

Table 1: Overview of corpus statistics (also is accessible at `https://www.orkg.org/orkg/comparison/R38012`). 'Total' and '%' columns show the number and percentage of instances annotated with the corresponding relation over all abstracts, respectively.

into two groups. One is based on quantitative analysis [27, 45]. In this group, researchers usually measure and compare the performance differences of the same NLP tool applied on the clean versus the OCR'd version of texts. The second group of studies is based on qualitative analysis. For example, [48] conducted a series of interviews with researchers to collect their feedback on using NLP techniques to analyze digital archives. Given the scholarly users' feedback, this study analyzed and summarized the impact of OCR errors on NLP techniques. In both groups of studies, a variety of NLP tasks were investigated, including tokenization, sentence segmentation, part-of-speech tagging, named entity recognition, topic modeling, document-level information retrieval, text classification, collocation, and authorial attribution. According to the findings of prior work, there is a consistent negative influence caused by uncorrected OCR on NLP tasks, some of which could even be "irredeemably harmed by OCR errors" [45].

In this study, we extend prior work by exploring the influence of OCR errors on relation classification from scholarly publications. Specifically, we aim to provide a systematic examination of BERT-based relation classifiers for their performances on both clean and OCR-noisy texts.

3 Corpora Preparation

The source data for this study includes two publicly available datasets [16, 30], each of which contains 500 born-digital scholarly abstracts with manual annotations of: (1) scientific terms; and, (2) semantic relation type of each related term pairs. To provide a comprehensive examination of BERT-based relation classifiers for recognizing scientific relations, from scholarly publications, with an emphasis on the real-world DL scenarios, we prepared both clean and noisy corpora. The details of each version of corpora are described below.

3.1 Clean Corpora

Regarding the clean corpora, in addition to directly using two selected raw datasets as two experimental corpora, we combined these two corpora into a third new corpus, which offers a more realistic evaluation setting because it provides a larger, more diverse task representation. Table 1 shows the statistics of our experimental corpora, each of which is detailed in the following subsections.

3.1.1 C1: The SemEval18 Corpus.

This corpus was created for the seventh Shared Task organized at SemEval-2018 [16]. The data was collected from scholarly publications in the ACL Anthology [1]: for a total of 500 abstracts, 350 of which were partitioned as training data and the remaining 150 as testing data. Originally, annotations in this corpus contained six discrete semantic relations that were defined to capture the predominant information content. Since the relation TOPIC has far fewer annotations than other types of relations, for our evaluation, we omit this relation type and consider the following five relation types: USAGE, RESULT, MODEL, PART-WHOLE, and COMPARISON.

3.1.2 C2: The SciERC Corpus

Although the second evaluation corpus SciERC [30] has the same number of annotated abstracts, unlike the SemEval18 corpus, this one contains diverse underlying data domains where the abstracts were taken from 12 artificial intelligence (AI) conference/workshop proceedings in five research areas: artificial intelligence, natural language processing, speech, machine learning, and computer vision. These abstracts were annotated with the following seven relations: COMPARE, PART-OF, CONJUNCTION, EVALUATE-FOR, FEATURE-OF, USED-FOR, and HYPONYM-OF. Similar to *C1*, this corpus was pre-partioned by the corpus creators. They adopted a 350/50/100 train/development/testing dataset split. Comparing *C2* with *C1*, we found that there are five relations, excepting CONJUNCTION and HYPONYM-OF, in *C2* that are semantically identical to the relations annotated in *C1*.

3.1.3 C3: The Combined Corpus

Finally, this evaluation corpus was created by merging *C1* and *C2*. In the merging process, we renamed some relations that are semantically identical but have different labels. First, USED-FOR in *C2* and USAGE in *C1* were unified as USAGE. Further, by observing relation annotations in *C1* and *C2*, we found that RESULT in *C1* and EVALUATE-FOR in *C2* essentially express a similar meaning but the arguments of these two relations were in reverse order. For example, "[accuracy] for [semantic classification]" is labeled as "accuracy" $\rightarrow$ EVALUATE-FOR $\rightarrow$ "semantic classification" in *C2*, which can be regarded as "semantic classification" $\rightarrow$ RESULT $\rightarrow$ "accuracy." Therefore, we renamed all instances annotated with relation EVALUATE-FOR in corpus *C2* into RESULT by flipping their argument order. By combining 1000 total abstracts with human annotations from two resources, our third evaluation corpus presents a comparatively more realistic evaluation scenario of large and heterogeneous data.

[1] https://aclanthology.org/

Clean: The *system* is based on a *multi-component architecture* where each component is responsible for identifying one class of unknown words .

Low Noise: The *syftcm* i^s based on a *multi-component d'architettura* whe^re eacb component is responsible for identifying one class of unknown vord3 .

High Noise: The *systcm* i^s based.1 on a *multi-component arcbitecture* wliere each.\u201d component i^s resj)onsible f^or identifying one.\u201d j.class of unknown.\u201d wordf .

Fig. 1: An illustrative example of a sentence in three versions: (1) clean; (2) low noise; and (3) high noise. The underlined phrases are human-annotated scientific terms in the original non-noisy dataset; the highlighted phrase in grey is the erroneous OCR introduced text.

3.2 Noisy Corpora with OCR Errors

Given that existing work on scientific relation classification focuses primarily on developing techniques using human-cleaned plain texts, so far as we know, there appears to be no prior work that has investigated the influence of OCR errors on scientific relation classification by utilizing NLP techniques. Given that, one major challenge for us to conduct this study is the lack of available data resources. Ideally, such datasets should satisfy two demands: (1) contain both clean and noisy versions of texts that are parallel in terms of content; and (2) have human-annotated relation types on clean data as ground truth.

Considering the high cost of directly preparing such a corpus in terms of both time consumption and human labor for tasks such as manual checking of text alignment and labeling relation types, we proposed an alternative strategy for this study that took advantage of the above off-the-shelf clean corpora with human annotations and replaced parts of their content with uncorrected OCR texts.

To employ this strategy, we adopted an existing token-level OCR-error dictionary [9] that lists a wide range of frequently uncorrected OCR'd tokens and their corresponding corrections. According to [9], these token pairs were collected from large-scale digitized English-language literature (178,381 volumes) published from 1700 to 1922. The total vocabulary size of this dictionary is 43,955. In order to provide a fine-grained analysis of the impact of OCR errors on BERT-based classifiers' performances with respect to the amount of text noise, rather than replacing all overlapping words between each scholar abstract and dictionary, we set a parameter to control the ratio of replacement. Moreover, considering the real-world uncorrected OCR in scanned texts usually has an irregular distribution, to simulate this pattern, for each scholarly abstract, we randomly selected a ratio value from a pre-set value group and our replacement was based on a random sampling of token candidates in the intersection of dictionary and the abstract content. In our empirical implementation, we prepared two levels of noisy corpus: (1) low noise corpus where the value of the replacement ratio per text is randomly chosen from the group {0.4, 0.5, 0.7}; and (2) high noise corpus where the ratio value per text is any of {0.9, 1.0}. Table 2 shows the distribution of OCR errors in each corpus. On average, corpus with low text noise has around 18% OCR errors in each text, while the corpus with high noise has around 49% errors per abstract.

	Low Noise	High Noise
SemEval18	18.09% $\pm$ 0.073	49.23% $\pm$ 0.103
SciERC	17.75% $\pm$ 0.070	48.88% $\pm$ 0.099
Combined	17.96% $\pm$ 0.071	48.97% $\pm$ 0.100

Table 2: Distribution of OCR errors in each noisy corpus

In order to have a better understanding of our prepared noisy corpora, we provide an illustrative example in Figure 1. As expected, sentences with high text noise have more OCR errors than ones with low noise. As we can see, text noise can vary even for the same words/phrases. For example, in the low noise sentence, the word "architecture" was replaced by "**d'**archi**tettura**", while in the high noise sentence, this word was changed to "arc**b**itecture". Given that many clean words in the dictionary have multiple versions of OCR errors, our random selection of one version for each word token's replacement is helpful in keeping the heterogeneous nature of OCR errors similar to a real-world scenario.

4 BERT-based Scientific Relation Classifiers

BERT [13] is a family of pre-trained language representations built on cutting-edge neural technology; it provides NLP practitioners with high-quality out-of-the-box language features that improve performance on many NLP tasks. These models return *contextualized* word embeddings that can be directly employed as features for downstream tasks. Further, with minimal task-specific extensions over the core BERT architecture, the embeddings can be fine-tuned to the task at hand with relatively little expense, in turn facilitating even greater boosts in task performance.

In this study, we employ BERT embeddings and fine-tune them with two classification strategies: (1) single-relation-at-a-time classification (SRC); and (2) multiple-relation-at-a-time classification (MRC). In the remainder of this section, we first describe the BERT models that we employ and then introduce our fine-tuned SRC and MRC classifiers, respectively.

4.1 Pre-trained BERT Variants

BERT models as pre-trained language representations are available in several variants depending on: (1) model configuration parameters such as model size and pre-training tasks; and, (2) pre-training data settings such as language, vocabulary case and text domain. In this study, we selected the following four core variants based on the combination of two key factors of the pre-training corpus, each with two categories: (1) text domain (i.e., generic or scientific); and (2) vocabulary case (i.e., cased or uncased).

BERT$_{BASE}$ [2] The first two models we use are in the category of pre-trained BERT$_{\text{BASE}}$. They were pre-trained on billions of words from the text data comprising the BooksCorpus (800M words) [55] and English Wikipedia (2,500M words). Our two selected models are: (1) a **cased** model (where the case of the underlying words were preserved when training BERT$_{\text{BASE}}$); and, (2) an **uncased** model (where the underlying words were all lowercased when training BERT$_{\text{BASE}}$).

SCIBERT [3] The next two models adopted in this study are in the category of pre-trained scientific BERT called SCIBERT. They are language models based on BERT but trained on a large corpus of scientific text. In particular, the pre-training corpus is a random sample of 1.14M papers from Semantic Scholar [3] consisting of the full text of published papers, 18% from the computer science domain and 82% from the broad biomedical domain. Like BERT$_{\text{BASE}}$, for SCIBERT, we use both its **cased** and **uncased** variants.

4.2 Fine-tuned BERT-based Classifiers

We implement the aforementioned BERT models within two neural system extensions that respectively adopt different classification strategies.

Single-relation-at-a-time Classification (SRC) Classification models built for SRC generally extend the core BERT architecture with one additional linear classification layer that has $K \times H$ dimensions, where K is the number of labels (i.e., relation types) and H denotes the dimension of the word embedding space. The label probabilities are further normalized by using a softmax function, and the classifier assigns the label with the maximum probability to each related concept pair.

[2] https://github.com/google-research/bert
[3] https://github.com/allenai/scibert

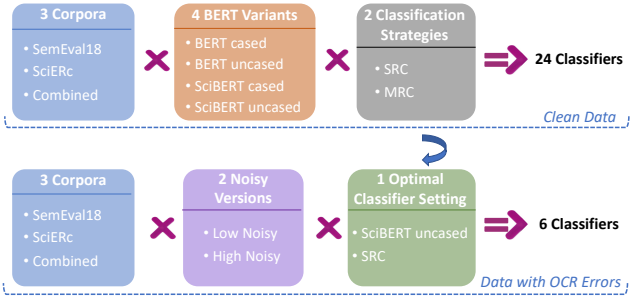

Fig. 2: Classifier statistics.

Multiple-relations-at-a-time Classification (MRC) This strategy is a more recent innovation on the classification problem in which the classifier can be trained with all the relation instances in a sentence at a time or predict all the instances in one pass, as opposed to separately for each instance. In this case, however, the core BERT architecture's self-attention mechanism is modified to efficiently consider the representations of the relative positions of scientific terms [50], which makes the encoding of the novel multiple-relations-at-a-time problem affordable. To obtain the classification probabilities, similar to the SRC strategy, the MRC is extended with a linear classification layer. However, at this time, this layer focuses on simultaneously calculating the probability per label assigned to each related term pair in a sentence. Finally, the label assignment per pair is the same as the one in SRC based on a softmax function.

5 Experiments

5.1 Experimental Setup

Figure 2 provides a brief summary of classifiers investigated in this study. In total, we built *24* classifiers on clean corpora and *6* classifiers on the corpora with OCR errors. Each corpus had been split into training/dev/testing set by the original dataset creators. To obtain the optimal classifiers on each corpus, we tuned the learning rate parameter η for values {2e-5, 3e-5, 5e-5}. For other parameters such as the number of epochs, we used default values in SCIBERT and BERT models.

With respect to the evaluation of classifiers' performances, we employed standard classification evaluation metrics including: Precision (P), Recall (R), F1-score ($F1$), and Accuracy (Acc).

	SRC						MRC						Avg±Std	
	SemEval18		SciERC		Combined		SemEval18		SciERC		Combined			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Bert-base uncased	76.42	71.74	84.6	77.25	81.75	77.38	80.4	79.98	83.42	74.84	80.84	76.29	81.24±2.84	76.25±2.78
Bert-base cased	73.58	71.14	85.32	77.92	78.73	74.38	79.55	78.44	83.72	75.07	79.42	74.8	80.05±4.14	75.29±2.65
Scibert cased	73.58	69.72	**86.86**	**79.65**	**84.46**	**81.60**	80.11	78.32	83.42	74.35	**81.80**	**77.68**	81.71±4.60	76.89±4.25
Scibert uncased	**80.97**	**79.42**	86.14	79.49	83.11	80.27	**81.82**	**80.54**	**84.33**	**77.44**	81.06	76.76	**82.91±2.04**	**78.99±1.54**
Avg. Scores	Acc. 84.10 F1 80.22						Acc. 82.35 F1 77.52							

Table 3: Scientific relation classification results over three datasets (SemEval18, SciERC, & Combined), four BERT model variants (BERT cased & uncased; SCIBERT cased & uncased), and two classification strategies (SRC & MRC). $Acc.$ is accuracy and $F1$ is the macro F1-score; Top scores are in bold.

5.2 Classification Results and Analysis

5.2.1 RQ1: What is the performance of eight BERT-based classifiers for scientific relation classification on clean data?

Table 3 provides an overview of classification results of eight classifiers trained and tested on the clean corpus based on either SRC and MRC classification strategy. Our comparison primarily focused on the following three key aspects of the classifiers.

The classification strategy, i.e., SRC vs. MRC. Given the Acc and $F1$ shown in Table 3, we observed that SRC outperformed MRC on two datasets except SemEval18. One characteristic of the SemEval18 dataset is that it has a significantly lower number of annotations than the other two datasets. Given that, we infer that the novel MRC strategy is more robust than SRC because its performance level was unaffected by a drop in the number of annotations.

Word embedding features, i.e., BERT vs. SCIBERT. Regarding the word embedding features encoded by different BERT-based models, SCIBERT outperformed BERT on all three corpora with higher accuracy and F1 scores. Since our experimental corpora are all scholarly data, as an expected result, word embeddings encoded by domain-specific BERT models can better capture the token-level semantic associations to support relation classification in the in-domain corpus than the embedding features encoded by the generic BERT models.

Vocabulary case in BERT models, i.e., cased vs. uncased. We observed that the uncased BERT models (SCIBERT: 82.91, BERT: 81.24) showed a higher classification accuracy than their cased counterparts (SCIBERT: 81.71, BERT: 80.05) on average. Further, the uncased models had an overall lower standard deviation in accuracy (SCIBERT: 2.04, BERT: 2.84) than the cased models (SCIBERT: 4.60, BERT: 4.14); comparisons on $F1$ are along similar lines. Hence, our results indicate that uncased BERT models can achieve more stable performances than cased variants.

In conclusion, with respect to the classification strategy, we observed that SRC outperformed MRC (see averaged scores in the last row in Table 3). Nevertheless, the advanced MRC strategy demonstrates consistently robust performance that remains relatively unaffected by the size of smaller datasets compared to the SRC (e.g. SRC vs. MRC results on the SemEval18 corpus). On the other hand, with respect to BERT word embedding variants, from the averaged scores in the last column in Table 3, the SCIBERT uncased model performed as the optimal model for encoding text features on scholarly articles. To further verify our findings, we conducted a set of statistical tests on the classification results. Considering that multiple datasets were employed in our examination and the distribution of the difference between any two samples' means may not be normally distributed, we decided to use Wilcoxon signed-rank test to explore the significance of prediction differences between any two variants of model setting. With a row-wise comparison of any two BERT word embedding variants' performance (based on F1 and Acc in Table 3 respectively) over three datasets across two classification strategies, we observed that uncased SciBERT-based classifiers significantly ($p<0.05$) outperformed cased and uncased BERT-based classifiers. To explore the performance difference between SRC and MRC, we conducted a column-wise comparison over three datasets. Similarly, this comparison was based on F1 and Acc respectively. The test results showed that there was a statistically significant ($p<0.05$) difference between SRC and MRC on the testing data from SciERC and Combined corpora. Although the classification results showed that MRC provided more benefits to BERT-based classifiers than SRC in SemEval18, such performance differences lacked statistical significance.

5.2.2 RQ2: Which of the seven studied scientific relation types were easy or challenging to be identified by the selected optimal classifiers?

Given the optimal classifier per classification strategy for each corpus (i.e., SCIBERT-based models, either cased or

Relationship Type SemEval18	SRC			MRC		
	P	R	F1	P	R	F1
USAGE	**87.22**	89.71	**88.45**	90.53	**87.43**	**88.95**
RESULT	78.26	**90.00**	83.72	**100.00**	75.00	85.71
COMPARE	85.71	85.71	85.71	75.00	85.71	80.00
MODEL-FEATURE	66.67	75.76	70.92	70.83	77.27	73.91
PART-WHOLE	79.25	60.00	68.29	70.83	72.86	71.83

Table 4: Per-relation classification results of the best BERT variant under SRC and MRC strategies on SemEval18.

Relationship Type SciERC	SRC			MRC		
	P	R	F1	P	R	F1
USED-FOR	**93.30**	91.37	**92.32**	**88.75**	90.24	**89.49**
CONJUNCTION	87.97	**95.12**	91.41	80.69	**95.12**	87.31
HYPONYM-OF	92.31	89.55	90.91	80.00	82.93	81.44
EVALUATE-FOR	82.29	86.81	84.49	84.44	83.52	83.98
COMPARE	72.73	84.21	78.05	83.87	68.42	75.36
PART-OF	66.04	55.56	60.34	65.52	60.32	62.81
FEATURE-OF	59.02	61.02	60.00	73.68	47.46	57.73

Table 5: Per-relation classification results of the best BERT variant under SRC and MRC strategies on SciERC.

Relationship Type Combined	SRC			MRC		
	P	R	F1	P	R	F1
CONJUNCTION	**92.56**	**91.06**	**91.80**	85.07	**92.68**	**88.72**
USAGE	91.30	88.98	90.13	**87.96**	87.71	87.84
HYPONYM-OF	89.39	88.06	88.72	83.12	78.05	80.50
COMPARE	86.89	89.83	88.33	73.85	81.36	77.41
RESULT	76.36	75.68	76.02	84.69	74.77	79.43
PART-OF	75.86	66.17	70.68	68.33	61.65	64.82
FEATURE-OF	58.02	75.20	65.51	60.28	68.00	63.91

Table 6: Per-relation classification results of the best BERT variant under SRC and MRC strategies on the Combined corpus.

uncased), we further examined these classifiers' ability to identify each type of relation labeled in the ground truth data. Tables 4, 5, and 6 show the results on SemEval18, SciERC, and Combined corpus, respectively.

Overview of relation type sensitivity. Overall, results in the three tables show that the USAGE (USED-FOR) relation is easier to identify than other relation types under both classification strategies. One possible explanation for this observation is that USAGE is the predominant type in all corpora, and therefore, classification models can readily learn the latent linguistic patterns of this relation type compared to the other types.

For challenging relations, we found that FEATURE-OF (MODEL-FEATURE) and PART-WHOLE (PART-OF) were more difficult to identify, with lower F1 scores compared with other relation types in all three tables. Our observations could be explained by two aspects. First, there is a high level of language expression diversity in these two types of relations, which poses a difficulty for the BERT-based classification models to capture the linguistic patterns of these two relation types. For example, by looking into instances labeled by PART-WHOLE (PART-OF), we found that the key signal of this relation type included varies forms such as ⟨A⟩ "is composed of" ⟨B⟩, ⟨A⟩ "...in" ⟨B⟩, and ⟨A⟩ ", a central instance of" ⟨B⟩. Moreover, the comparatively lower number of annotations of these two relation types in the corpora, especially for the SCIERC corpus, decreases the generalizability of models when predicting these two relation types, and thus leads to low F1 scores on unseen testing examples.

Impact of classification strategies on per-relation type classification. By looking into classification strategies for the SemEval18 corpus (see Table 4), we found that the SRC classifier and MRC classifier obtained the same classification rank order for USAGE, MODEL-FEATURE and PART-WHOLE, but the opposite order for RESULT and COMPARE. In particular, the SRC classifier performed better at identifying COMPARE, while the MRC classifier was able to better recognize RESULT. As to SCIERC (see Table 5), notably, the ability of classifying HYPONYM-OF dropped significantly from the SRC to the MRC strategy, suggesting that the linguistic pattern of HYPONYM-OF is hard to be captured when this relation type is mixed with other types together.

Impact of heterogeneous human annotations on per-relation type classification. The combination of SemEval18 and SCIERC as the Combined corpus makes this corpus more heterogeneous and realistic. Given that, we further explored whether the mix of the two groups of human annotations from different corpus sources showed any influence on the classifier's ability to identify each relation type.

Interestingly, differing from the results shown in Table 4 (SemEval18) and Table 5 (SCIERC) where the F1 score of predicting USAGE (USED-FOR) ranked first in both SRC and MRC strategies, we observed that CONJUNCTION was the easiest relation type to classify in the Combined corpus for both SRC and MRC classifiers (see Table 6). Moreover, the F1 score of CONJUNCTION in Table 5 is higher than the one in Table 6. Since there is no CONJUNCTION relation in the SemEval18 corpus, we can infer that testing examples of this relation type should be the same as the ones in the SCIERC corpus. Given that, our results suggest that there may exist inconsistencies in the two groups of human annotations of the relation USAGE, which influenced the model's learnability of this relation type. On the other hand, the uniform annotation of CONJUNCTION from the SciERC corpus could benefit the model's ability to capture the consistent latent linguistic pattern for this relation type in the Combined

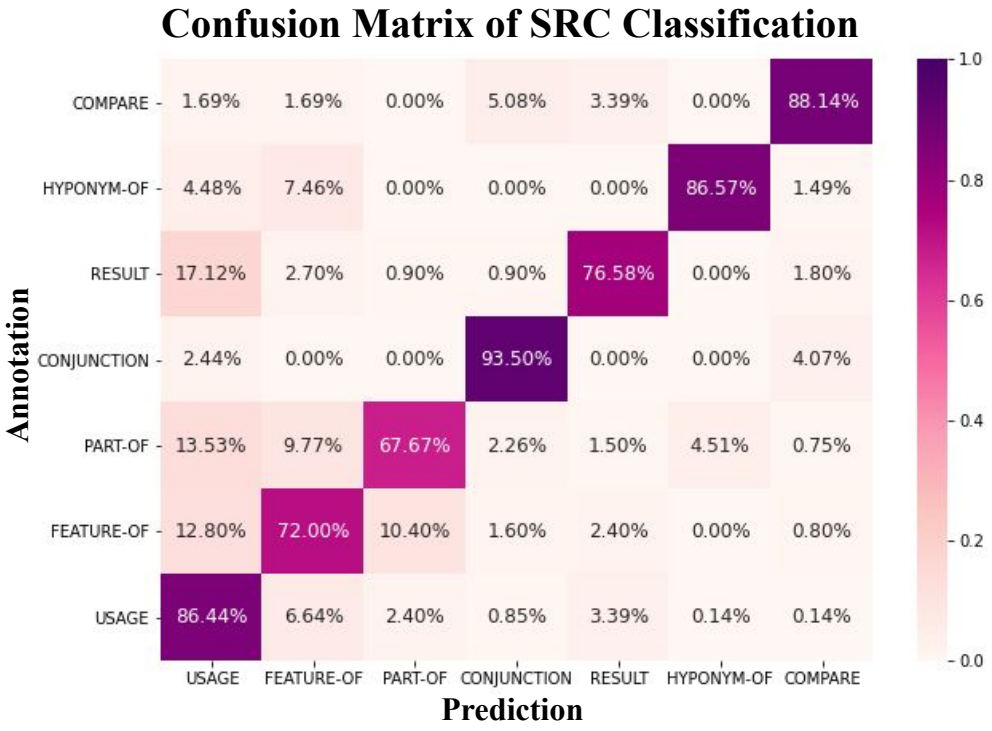

Fig. 3: The confusion matrix of SRC classification on the Combined corpus.

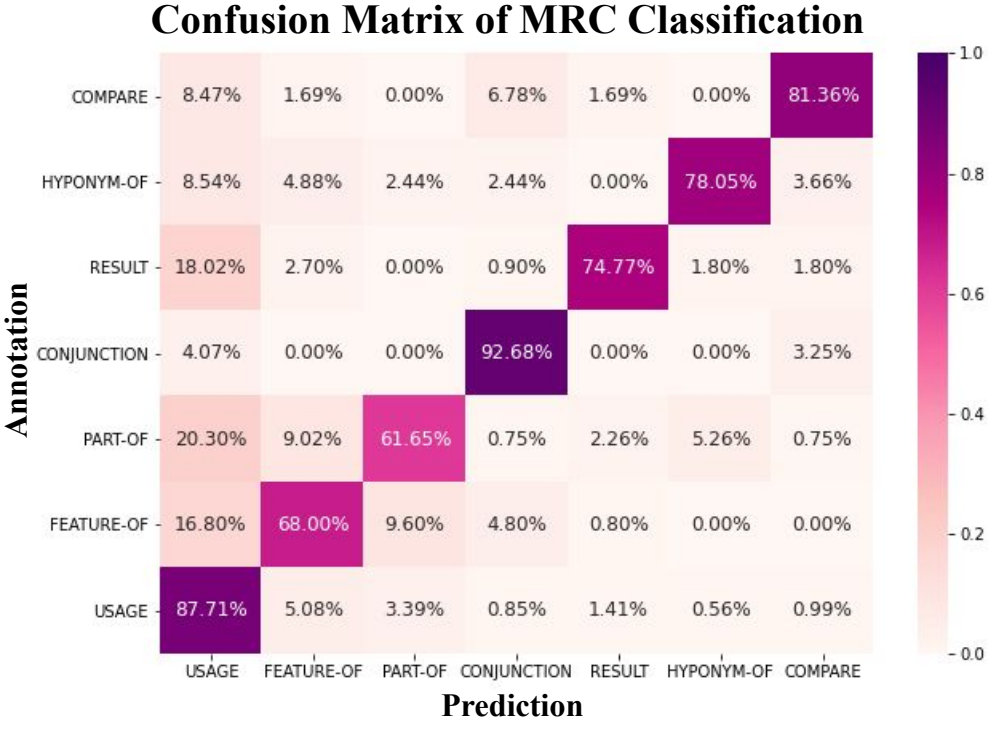

Fig. 4: The confusion matrix of MRC classification on the Combined dataset.

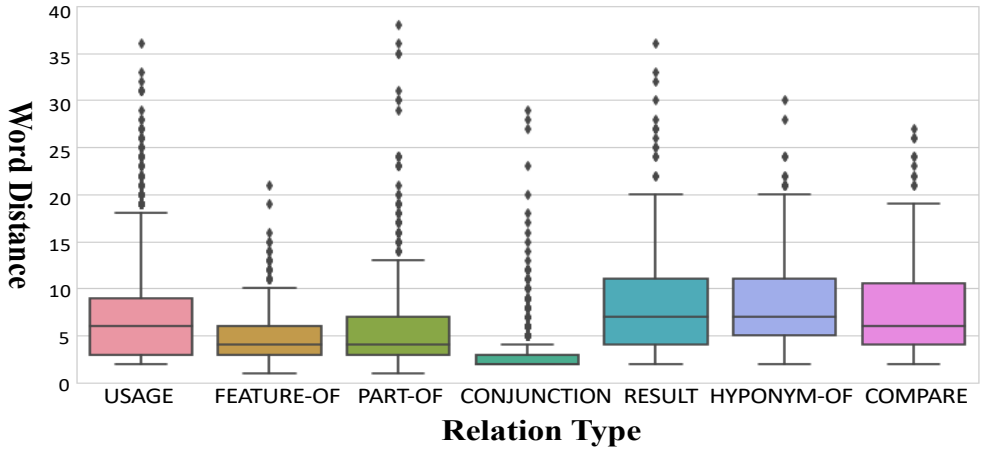

Fig. 5: Distributions of word distances between scientific term pairs in abstracts in combined corpus.

corpus, and hence, allowed it to be classified more effectively. Nevertheless, we can see that the classification F1 for USAGE was nearly the average of the scores from each individual corpus respectively. This offers a potential indicator of expected classification behavior on a heterogeneous corpus from more than one source.

5.2.3 RQ3: What kinds of prediction errors are frequently made?

Confusion matrix of SRC and MRC classification. A closer look at misclassifications in the Combined corpus is depicted in the confusion matrices in Figure 3 for the SRC classifier and Figure 4 for the MRC classifier. Results in both figures show that classifiers tend to predominantly misclassify other relation types such as USAGE (e.g., RESULT, PART-OF, and FEATURE-OF), especially in the MRC classification. One possible reason is that the USAGE relation has the largest number of annotated instances, which caused the classifiers to be biased to this relation type. In general, unbalanced distribution of training samples (see the details in Section 3) is one of the main causes in confusion learned in machine learning systems. At the same time, the strategy of making predictions on multiple relation types simultaneously can boost such biases to learning models.

In particular, FEATURE-OF, PART-OF, and USAGE are highly likely to be confused with each other in both classification strategies, suggesting that these three relationships have closer semantic associations than other relation types, and therefore making it difficult for the machine to differentiate between these three relation types. We further verified this finding by referring to the annotation guidelines of SemEval18 and SCIERC. For example, in SCIERC's guideline, the instance ⟨B "belongs to" A⟩ is semantically similar with the instance ⟨B " is a part of" A⟩. However, the former relationship was defined as FEATURE-OF, while the latter one was a case of PART-OF. We also found that HYPONYM-OF (around 6%) is more likely to be misclassified as FEATURE-OF (in addition to USAGE), but not vice versa. This one-sided relationship loosely demonstrates that the machine might learn a relation hierarchy between these two relation types, i.e. HYPONYM-OF subsumes FEATURE-OF, but not the other way around.

In conclusion, our findings show that errors in the classification of scientific relations by the optimal BERT-based classifier can be caused by three factors: (1) unbalanced data distribution; (2) semantic ambiguity of pre-defined relation types; and, (3) hierarchical semantics of pre-defined relation types.

Word distance distribution. To offer another pertinent angle on the classifier error analysis, we computes the distribution of word distances between related scientific term pairs for each type of relation on the Combined corpus. The result is depicted in Figure 5. In general, the majority of box plots shown in Figure 5 are skewed with a long upper whisker and a short lower whisker. This pattern indicates that the distance between paired scientific terms is typically closed in the text. As opposed to other relations, the word distance of CONJUNCTION is much shorter, which makes sense because term pairs with this relationship are typically connected by a single connection term such as "and" and "or". This consistent pattern could be another reason why CONJUNCTION is comparatively easier to be classified than other relations. Further, the average word distance of FEATURE-OF, PART-OF, HYPONYM-OF, and COMPARE is closer to the lower quartile than the other relations. Such varied distribution may bring challenges for a classifier to identify these relations. Notably, the similar median value and spread range between FEATURE-OF and PART-OF could account for why they are challenging for the classification models to identify.

5.2.4 RQ4: How do OCR errors impact the overall optimal classifier's performances?

Following the analysis of classification results on clean corpora presented in the earlier sections, we found that the uncased SciBERT built under SRC was optimal to identify scientific relations in general. In order to provide comparable results across corpora, in this section, we applied this optimal model and further analyzed its classification performance on the noisy version of all evaluation corpora. Our goal was to explore if there is any consistent pattern of the impact of OCR noise on predictions.

Tables 7, 8, and 9 show the classification results on SemEval18, SciERC, and Combined corpus, respectively. Our analysis primarily focused on two aspects: (1) the impact of OCR errors from various combinations of noisy training and testing splits in classifications; and, (2) other potential data factors, such as corpus size, associated with the impact of OCR errors. The details of each aspect is described below.

Impact of OCR noise in testing data. Results in three tables showed that the growth of OCR errors in testing data led to a rapid drop off in the classifier's predictability performance regardless of the amount of text noise in the training data. For example, the classifier trained on the clean texts per corpus (see Table 7/ 8/ 9) had a loss of F1 scores at around 5% for the testing data with a low amount of OCR errors, and around 20% for the one with high a high amount of text noise. Given the same training data, our observation shows that the loss of predictions increases in accordance with the growth of OCR errors in texts.

Impact of OCR noise in training data. Regarding the number of OCR errors in the training data, in each corpus we observed that the increasing of text noise in training data helped the classifier in improving its robustness, with a lower loss of F1 score, when making predictions on testing examples, especially for the training set with a high amount of text noise. As shown in the three tables, classifiers trained on the clean data had a performance loss of around 20% F1, while for the models built upon high noise training data, the loss decreased to 6% F1 score on average. Our observations indicate that OCR errors in texts have an obvious impact on the performance of BERT-based classifier for identifying scientific relations within sentences. Interestingly, text noise in the training data has a regularization effect on the transformer-based neural network model during its learning process, which essentially benefits the model to improve its generalization ability to process the noisy unseen data.

Impact of the size of OCR noise among corpora. By further comparing the classification results among three corpora, we found that the predictability of classifiers on the SemEval18 corpus was much worse, with lower F1 scores than the classification models' performances on the other two corpora when the training data is noisy. For example, the predication difference could be around 10% for each type of testing set when the training data is high noise. Given that the SemEval18 corpus has a smaller corpus size than the other two corpora, our observation suggests that large corpus size could alleviate the impact of text noise on BERT-based classification models for identifying scientific relations.

Training / Testing	Clean		Low Noise		High Noise	
	F1	Loss	F1	Loss	F1	Loss
Clean	**79.42**	-	**75.11**	-	**64.82**	-
Low Noise	73.16	5.26↓	72.21	2.90↓	62.72	2.10↓
High Noise	56.35	23.07↓	58.35	16.76↓	58.26	6.56↓

Table 7: Classification results of uncased SciBERT on noisy SemEval18 corpus in SRC. Top F1 scores are in bold.

Training / Testing	Clean		Low Noise		High Noise	
	F1	Loss	F1	Loss	F1	Loss
Clean	**79.49**	-	77.51	-	**78.05**	-
Low Noise	75.16	4.33↓	**77.92**	-	76.35	1.70↓
High Noise	60.11	19.38↓	69.94	7.57↓	72.11	5.94↓

Table 8: Classification results of uncased SciBERT on noisy SciERC corpus in SRC. Top F1 scores are in bold.

Training / Testing	Clean		Low Noise		High Noise	
	F1	Loss	F1	Loss	F1	Loss
Clean	**80.27**	-	**78.64**	-	**77.10**	-
Low Noise	72.84	7.43↓	77.41	1.23↓	74.94	2.16↓
High Noise	57.43	22.84↓	69.29	9.35↓	70.93	6.17↓

Table 9: Classification results of uncased SciBERT on noisy Combined corpus in SRC. Top F1 scores are in bold.

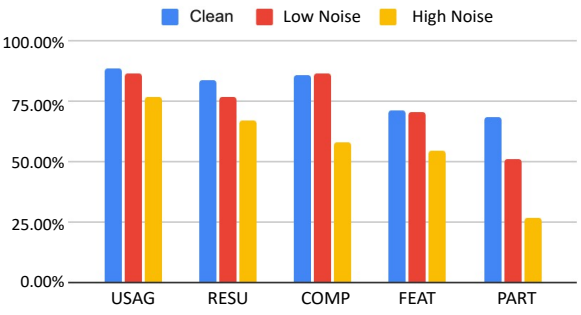

Fig. 6: Per-relation type classification results on clean, low-noise and high-noise SemEval18 corpus.

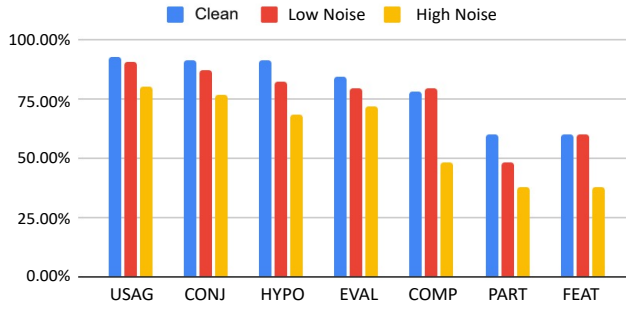

Fig. 7: Per-relation type classification results on clean, low-noise and high-noise SciERC corpus.

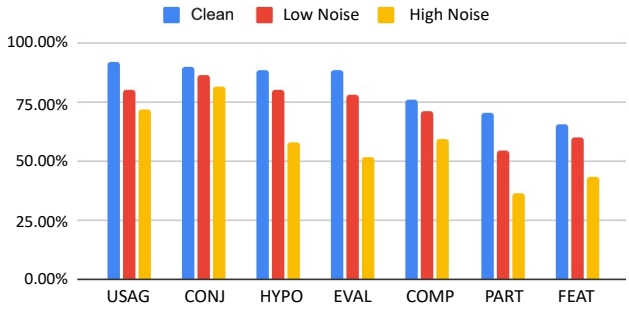

Fig. 8: Per-relation type classification results on clean, low-noise and high-noise on Combined corpus.

5.2.5 RQ5: Which of the seven studied relation types are more robust to OCR errors?

Given that the majority of state-of-the-art scientific classification techniques are developed using clean corpus, to investigate the vulnerability of such techniques to text noise with respect to each type of relations, we examined the ability of our optimal classifiers trained on the clean corpus to identify each relation type from texts with or without OCR errors. Figures 6, 7, and 8 show the results on three corpora respectively.

In general, results for three corpora showed that the difficulty of the classifier to identify each relation type tends to be consistent regardless of the amount of noise in the testing texts. Comparatively, COMPARE, PART-OF (PART-WHOLE), and HYPONYM-OF are three top relation types for which the classifier's performances were more sensitive to the large amount of OCR errors in terms of a high loss of F1 scores compared with other relation types, which suggests that these three types of relationships between scientific concepts are easily broken by text noise. In contrast, the predictability of each classifier on USAGE and RESULT (EVALUATE-FOR) is more robust against text noise than on other relation types. Given that, we infer that there might exist a strong semantic association between USAGE and RESULT relations, which could overcome the disturbance of OCR errors to some extent.

By looking further into the influence of text noise by OCR error amount, we found that the difference in the classifier's ability to identify most relation types between the clean and low-noise texts is much smaller than the difference between low versus high noise texts. The only exception lies in the PART-OF (PART-WHOLE) relation. This result indicates that our optimal classifier can be robust to a small ratio of text noise (i.e., around 10%) when making predictions on the majority of pre-defined relations except PART-OF (PART-WHOLE).

5.3 Use Case of Scientific Relation Classification for Scholarly Knowledge Graph Construction

As a practical illustration of the relation triples studied in this work, finally, we built a knowledge graph from the human annotations in the Combined corpus. Figure 9 provides an illustrative example of the visualization of the resulting knowledge graph, which includes: (1) a graph at corpus-level (shown in the upper right); and, (2) two graphs at concept-level, e.g. ego-network for the concept node "machine translation" (shown in the upper left) and "words" (shown at the bottom).

Looking at the corpus-level graph, we observed that generic scientific terms such as "method," "approach," and "system" were the most densely connected nodes, as expected since these generic terms are usually found across research areas. In the zoomed-in ego-network of "machine translation," we can see that HYPONYM-OF is meaningfully highlighted by its role linking "machine translation" and its sibling nodes (i.e. other research tasks) including "speech recognition," and "natural language generation" to the parent node "NLP problems." The concept "lexicon" is usually used in (i.e., USAGE) research on "machine translation" and "operational foreign language." The CONJUNCTION link connects the term "machine translation" and "speech recognition," both of which aim at translating information from one source to the other one. Regarding the ego-network of "words", we found that this term mainly involve two types of semantic relations: (1) being PART-OF other natural language components such as "corpus", "song_lyrics", and "sentences"; or, (2) working as features (i.e., FEATURE-OF) in NLP computations such as obtaining "part_of_speech_information" and capturing word-level "co-occurrence_patterns".

In addition to looking singly into the semantic association across the node as well as into link representations in the ego-network, and by further considering the network structure, we can infer the semantic hierarchy of some associated scientific concepts. For example, both "machine translation" are connected with "ranking tasks" and "NLP problems" by the relation HYPONYM-OF, and "ranking tasks" is used for solving (i.e., USAGE) "NLP problems." This triangle structure suggests that "ranking tasks" is a more fine-

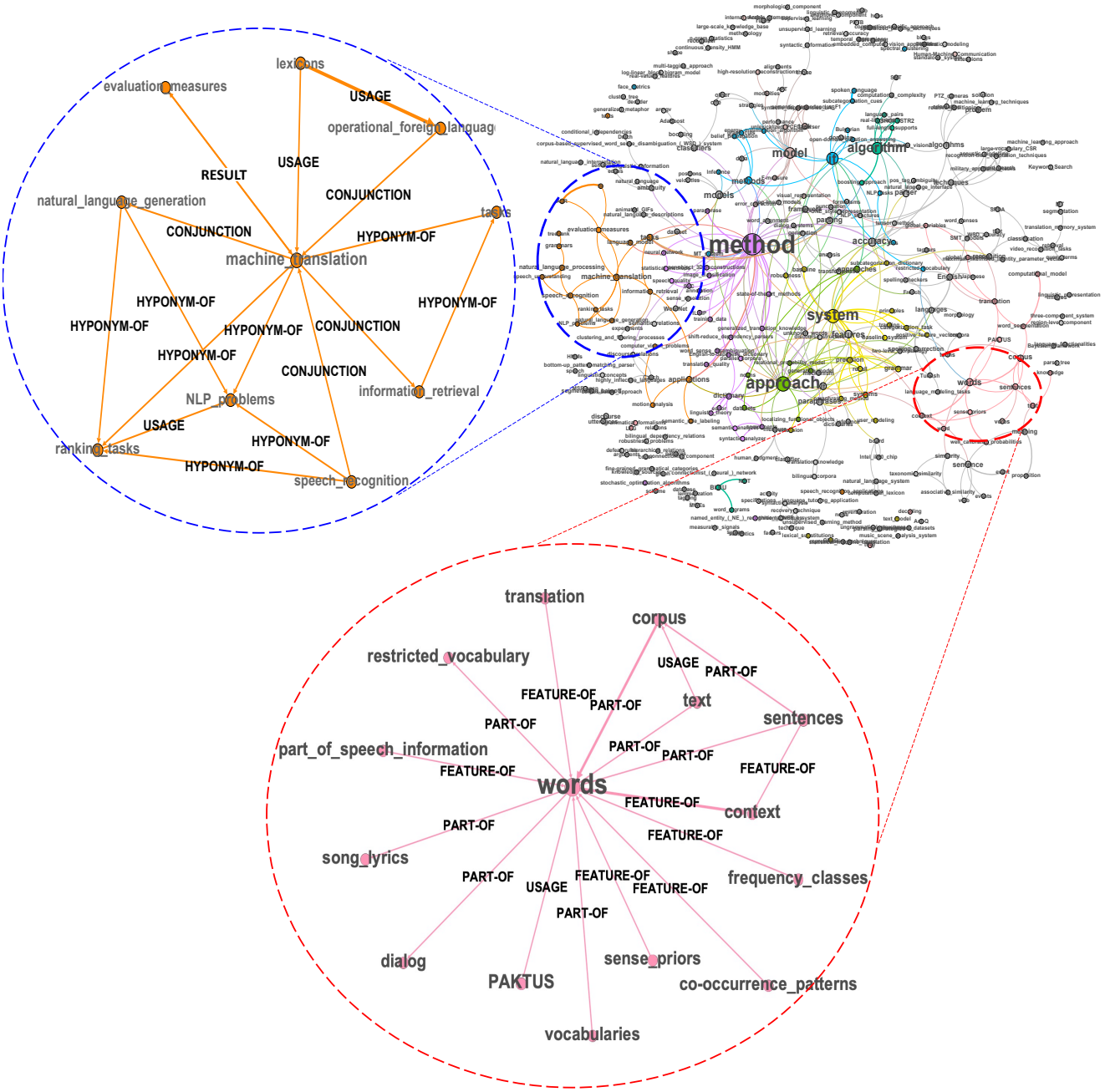

Fig. 9: A knowledge graph constructed from the relation triples in the Combined corpus. Ego-network for the term "machine_translation" and "words" are presented as two illustrative examples. The node size is determined by node weighted degree. Colors denote the modularity classes based on the graph structure. The graph was generated using Gephi (https://gephi.org/)

grained hypernym to "machine translation" than the term "NLP problems." Additionally, by comparing the structure of the ego-network of "machine-translation" with the one of "words", we observed the "machine-translation"-centered graph is more dense than the "words"-centered graph. In essence, this observation indicates that the term "words" is a more generic concept than "machine_translation" because it tends to be associated with diverse independent concepts.

In summary, given the pair of related scientific concepts with the identified relation type, we can construct a knowledge graph to represent the semantic association among scientific concepts either at macro-level in terms of the whole corpus or at micro-level with regards to the fine-grained connections related to a specific concept.

6 Discussion

6.1 BERT-based Model Recommendations for Scientific Relation Classification

Based on the findings in this study, we provide the following recommendations to stakeholders of digital libraries for applying the optimal technique to automatically classify scientific relations from scholarly articles:

- Compared with the generic domain, using corpus that specifically consists of scholarly publications to pre-train BERT models benefits building classifiers in recognizing the relationship between scientific concepts.
- With respect to the classification strategy, SRC outperforms MRC in general.
- Overall, uncased BERT models outperform cased ones in terms of higher accuracies and more stable performances in predictions.
- Given noisy texts containing OCR errors, if the amount of text noise is small, such as around 10%-20%, the optimal classification model (i.e. built with uncased SCIBERT in SRC classification) trained on the clean corpus could be robust against this text noise when making predictions.
- For noisy texts with an uncertain amount of OCR errors, employing noisy data to train the classifiers would be helpful in building the models' generalization ability to process the texts regardless of the amount of OCR errors.
- In SCIBERT-based SRC classification, the large corpus size with abundant relation annotations is helpful in: (1) improving the prediction accuracy; and (2) alleviating the influence of text noise in the training data for building a classifier.
- Regarding relation type annotations, the large number of annotations for a relation type in the training set can help the classification model to improve its learnability on this relation type.
- For each pre-defined relation, the fixed syntactic structure in expressions benefits the classifier in discriminating between different relation types.
- Text noise is less influential on the classifier when predicting concrete scientific relations such as USAGE and RESULT (EVALUATE-FOR) compared with other relations, while this situation is contrary for some relations indicating concept hierarchies such as COMPARE and HYPONYM-OF.

6.2 Study Limitations and Challenges

With a systematic review of our evaluation process, several research challenges and resulting study limitations arise, which can be summarized in three aspects: (1) human annotation collection; (2) noisy corpus preparation; and (3) evaluation methods. Each aspect is described as follows.

6.2.1 Human Annotation Collection

Cost and resource of annotations As mentioned in prior work [16, 30], one common challenge in research on scholarly information extraction (including both entity and relation extraction from scholarly records) is the high cost and limited resources for collecting human annotations as ground truth for training and/or testing learning-based methods. This is mainly because annotations on scientific information not only cost time and human labor similar to that involved in information extraction tasks in generic domains (e.g., newswire), they also require a higher level of annotators' expertise in the specific scientific domain of the target scholarly articles, for which it is comparatively more difficult to find adequate annotators. In addition, the expense of such annotations should be high. Due to these challenges, the number of human annotations on scientific information is usually limited, and the number of research areas involved by existing accessible human annotations is barely sufficient.

Influenced by the aforementioned issues, our evaluation mainly focused on the corpora covering AI-related research areas, which potentially leads to some uncertainties in the performance of BERT-based classifiers for identifying scientific relations in other research areas such as biology and physics.

Quality of annotations as ground truth Following our findings, the ambiguity of relation type definitions is one of the key factors leading to the classifiers' misclassifications. There are two main aspects contributing to such ambiguity. First, the meaning of some relation types' definitions makes it difficult for humans to have a clear understanding of these relation types when annotating the data. For example, in the annotation guidelines of SCIERC dataset [30], ⟨B⟩ "belongs to" ⟨A⟩ was defined as an instance of FEATURE-OF, while ⟨B⟩ "is a part of " ⟨A⟩ was defined as a case of PART-OF. In essence, both instances are quite similar in terms of the semantic meaning. Second, different annotation systems may lead to inconsistent human annotations. For example, for two relations having a hierarchical semantic association, i.e., HYPONYM-OF and PART-OF, the guideline of SEMEVAL18 dataset only considers PART-OF, while SCIERC's guidelines ask the annotators to label both relation types. Given this guideline difference, annotators are likely to annotate the real HYPONYM-OF relationship in the SEMEVAL18 dataset by the tag PART-OF.

In addition to annotation ambiguity, there also exists annotation biases in the corpora, such as the unbalanced distribution of relation labels, which can lead to the preference of classifiers to recognize some well-represented relations(e.g.,

USAGE). Due to these challenges leading to the uncertainty of annotation quality, there may exist a potential risk of underestimating the ability of BERT-based classifiers for identifying scientific relations in our evaluations.

6.2.2 Digitization-based Noisy Corpus Preparation

In order to investigate the impact of text noise caused by digitization on relation classification techniques, building an ideal noisy corpus typically requires two elements: (1) the corpus should have both clean and real-world OCR'd texts that are parallel in terms of content; and, (2) the clean corpus should have a high quality. However, it is not always possible to satisfy these two demands. One major reason is that the original version of digitized scholarly records were usually published in printed pages, which means such resources lack an off-the-shelf electronic version of texts. Given that, getting access to the fully clean texts of these scholarly resources is difficult. Besides, the aforementioned challenges of human annotation also exist in this data preparation process.

Although our presented strategy for constructing a noisy corpus in this paper can be an alternative method to address the above challenges, adding common word-level OCR noise into the clean texts based on a dictionary may not fully reflect the data patterns of the real-word digitized library collections. This limitation might involve the artifacts of engineering in our noisy corpus for further investigations.

6.2.3 Evaluation Methods

Following the standard practice of prior work [50,8], we used existing popular pre-trained BERT models as off-the-shelf tools and further fined-tuned these models with each of our corpora to build classifiers for scientific relation classification. Our investigation primarily concentrated on the benefits of fine-tuned BERT-based models brought to the classifiers. There exist some open questions regarding the characteristics of BERT's or SciBERT's pre-training settings and their influences on the downstream application. Example questions could be: (1) how is the quality of the corpus used to pre-train BERT; and, (2) how does the influence of this factor on BERT-based classifiers for identifying scientific relations? These issues are worthwhile to be explored in future. In addition, in this study, we mainly focused on the corpora covering AI-related research areas. The assessment of classifiers for identifying scientific relations in other domains need to be further studied.

6.3 Potential Future Work

To further assist digital library designers and librarians who want to build structural semantic representations over scholarly articles using scientific relation classifiers, there are three main avenues that are worthy of future exploration.

6.3.1 Rethinking Human Annotation Guidelines

Challenges in guaranteeing the quality of human annotations show that the process of defining scientific concepts and their relationships is still an open question, which can be further explored by experts. Researchers who are interested in this field can conduct their following studies in two directions.

On the one hand, a formal understanding of the semantic relationships among terms in a specific research domain is critical to improve the clarity of annotation guidelines on this domain's scholarly information extraction. In particular, we suggest the consideration of semantic relation types should not only contain the semantic associations between scholarly terms, but also require the hierarchical structure of relation types.

On the other hand, it would be helpful if the strategy for designing annotation guidelines could follow the ultimate application goal. In practice, building a scholarly knowledge graph can be used for organizing general scholarly knowledge within a single or across multiple research domain(s) or managing knowledge with a specific focus, such as the evolution of scientific works' contributions. With different applications of scholarly knowledge graphs, the corresponding annotation rules on scientific terms and pre-defined relation types might be different. For example, annotations on the general scientific concepts shown in Figure 9 such as "method", "approach", "system", and "algorithm" are limited to informativeness for indicating the specific contributions of each research work.

6.3.2 Benchmark Corpus from Multiple Domains

While scholarly records in digital libraries usually cover various domains, publicly accessible corpora used for developing scholarly information extraction techniques primarily in several specific domains such as biology [29,35] and computer science [31,16]. Given that scholarly publications in different domains may cover various relation types, and increasingly, language expressions such as word choice and writing style could be different, there is a demand for building a benchmark corpus consisting of scholarly records in a wide range of domains, which can be helpful in providing a comprehensive evaluation for the state-of-the-art relation classification techniques developed for extracting scholarly semantic information.

6.3.3 Evaluation on Open Information Extraction Techniques

In addition to building classifiers to identify pre-defined relations, techniques that are developed under the paradigm of open information extraction to identify more diverse relational triples "without requiring any relation-specific human input" [7] can an alternative yet promising strategy for extracting scholarly information, especially for identifying scientific relational tuples from scholarly records in various domains with massive uncertain semantic relations in advance. The further examination of state-of-the-art techniques in this field for scientific relation identification and the trade-offs between such techniques with BERT-based classification on pre-defined relations could be a valuable avenue to pursue.

7 Conclusions

We have investigated the scientific relation classification task to support building scholarly knowledge graphs based on digital library collections. We provide a comprehensive view of eight BERT-based classification models on three clean corpora, which differ usefully in terms of corpus size and annotation guidelines. Moreover, considering many scholarly records in real-world digital libraries are digitized with OCR errors, we further prepared three noisy corpora corresponding to the clean ones and investigated the effect of OCR errors on the optimal BERT-based classifier identified from clean data. The presented empirical study in this paper contributes to the digital library stakeholder's understanding of state-of-the-art NLP techniques for identifying semantic relations from scholarly publications, which can provide practical benefits for identifying the optimal NLP tool to build scholarly knowledge graphs of digital library collections.

Our observations indicate that the performance of classifiers on clean texts is mainly associated with two aspects. First, from the perspective of training algorithms, three main factors, including classification strategies, the pre-training corpus domain and vocabulary case, determine the optimal model to apply in practice. Second, with respect to the annotation of scientific relations for training, there are two key factors that influence the ability of a BERT-based classification model to identify each relation type: (1) the number of annotations of each relation type; and, (2) the regularity of each relation's syntactic context. With further exploration on OCR noise impacts, we found that text noise caused by digitization has an obvious negative influence on the performance of BERT-based classifiers when identifying scientific relations, especially when the ratio of noise is high (e.g., 49%). Comparatively, relations with more concrete semantic meaning such as USAGE and RESULT are more beneficial with the classifiers' robustness to OCR noise than other relations, while relations showing concept hierarchy like COMPARE and HYPONYM-OF are more likely to stimulate the vulnerability of classifiers to the noise.

The overall insights in this study suggest that the uncased SCIBERT-based classification model built under SRC strategy is the optimal choice for scientific relation classification in general. Regarding the corpus with OCR noise, we suggest DL stakeholders employ noisy data to build classifiers because the heterogeneous nature of OCR noise in training data is helpful with the generalization ability of classification models for processing unseen data.

Acknowledgements The authors would like to thank anonymous reviewers for their constructive comments on this paper. We also appreciate Alaine Martaus, Jacob Jett and Yuerong Hu from University of Illinois at Urbana-Champaign for their helpful paper proofreading.

References

1. Pubmed. https://pubmed.ncbi.nlm.nih.gov/
2. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of the 5th ACM Conference on Digital Libraries. pp. 85–94 (2000)
3. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al.: Construction of the literature graph in semantic scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). pp. 84–91 (2018)
4. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. pp. 1–6 (2018)
5. Auer, S., Mann, S.: Toward an open knowledge research graph. The Serials Librarian **76**, 1–7 (2019)
6. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 546–555 (2017)
7. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 28–36 (2008)
8. Beltagy, I., Lo, K., Cohan, A.: 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. ACL, Hong Kong, China (Nov 2019)
9. Center, H.R.: Genre-specific word counts for 178,381 volumes from the hathitrust digital library [v.0.1] (2015), https://wiki.htrc.illinois.edu/display/COM/Word+Frequencies+in+English-Language+Literature%2C+1700-1922
10. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: Towards a stratified learning approach to predict future citation counts. In: IEEE/ACM Joint Conference on Digital Libraries. pp. 351–360. IEEE (2014)

11. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 423–429. ACL (2004)
12. Davletov, F., Aydin, A.S., Cakmak, A.: High impact academic paper prediction using temporal and topological features. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. pp. 491–498 (2014)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. ACL, Minneapolis, Minnesota (Jun 2019)
14. Esakov, J., Lopresti, D.P., Sandberg, J.S.: Classification and distribution of optical character recognition errors. In: Document Recognition. vol. 2181, pp. 204–216. International Society for Optics and Photonics (1994)
15. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. vol. 32, pp. 5779–5786 (2018)
16. Gábor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.: SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 679–688 (2018)
17. Gao, T., Han, X., Liu, Z., Sun, M.: Hybrid attention-based prototypical networks for noisy few-shot relation classification. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. vol. 33, pp. 6407–6414 (2019)
18. Gonçalves, G.D., Figueiredo, F., Almeida, J.M., Gonçalves, M.A.: Characterizing scholar popularity: A case study in the computer science research community. In: IEEE/ACM Joint Conference on Digital Libraries. pp. 57–66. IEEE (2014)
19. Hallo, M., Luján-Mora, S., Maté, A., Trujillo, J.: Current state of linked data in digital libraries. Journal of Information Science **42**(2), 117–127 (2016)
20. Hashemi, S.H., Neshati, M., Beigy, H.: Expertise retrieval in bibliographic network: A topic dominance learning approach. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 1117–1126 (2013)
21. Haslhofer, B., Isaac, A., Simon, R.: Knowledge graphs in the libraries and digital humanities domain. Sakr S., Zomaya A. (eds) Encyclopedia of Big Data Technologies. pp. 1–8 (2018)
22. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 541–550 (2011)
23. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. pp. 243–246. ACM, New York, NY, USA (2019)
24. Jett, J., Cole, T.W., Han, M.J.K., Szylowicz, C.: Linked open data (lod) for library special collections. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 1–2. IEEE (2017)
25. Jia, W., Dai, D., Xiao, X., Wu, H.: Arnor: Attention regularization based noise reduction for distant supervision relation classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1399–1408 (2019)
26. Jiang, M., Diesner, J.: A constituency parsing tree based method for relation extraction from abstracts of scholarly publications. In: Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13). pp. 186–191 (2019)
27. Jiang, M., Hu, Y., Worthey, G., Dubnicek, R.C., Capitanu, B., Kudeki, D., Downie, J.S., et al.: The Gutenberg-HathiTrust parallel corpus: A real-world dataset for noise investigation in uncorrected OCR texts. Poster at iConference 2021 (2021), `http://hdl.handle.net/2142/109695`
28. Klampfl, S., Kern, R.: An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles. In: International Conference on Theory and Practice of Digital Libraries. pp. 144–155. Springer (2013)
29. Kruiper, R., Vincent, J.F., Chen-Burger, J., Desmulliez, M.P., Konstas, I.: A scientific information extraction dataset for nature inspired engineering. arXiv preprint arXiv:2005.07753 (2020)
30. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3219–3232 (2018)
31. Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., Hajishirzi, H.: A general framework for information extraction using dynamic span graphs. 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. 3036–3046 (Jun 2019)
32. Manning, C.D.: Computational linguistics and deep learning. Computational Linguistics **41**(4), 701–707 (2015)
33. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011 (2009)
34. Mohapatra, D., Maiti, A., Bhatia, S., Chakraborty, T.: Go wide, go deep: Quantifying the impact of scientific papers through influence dispersion trees. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 305–314. IEEE (2019)
35. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: A survey. Scientometrics **117**(3), 1931–1990 (2018)
36. Penfield, T., Baker, M.J., Scoble, R., Wykes, M.C.: Assessment, evaluations, and definitions of research impact: A review. Research Evaluation **23**(1), 21–32 (2014)
37. Pradhan, T., Pal, S.: A multi-level fusion based decision support system for academic collaborator recommendation. Knowledge-Based Systems **197**, 1–23 (2020)
38. Quan, T.T., Hui, S.C., Fong, A.C.M., Cao, T.H.: Automatic generation of ontology for scholarly semantic web. In: International Semantic Web Conference. pp. 726–740. Springer (2004)
39. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 148–163. Springer (2010)
40. Saggion, H., Ronzano, F.: Scholarly data mining: making sense of scientific literature. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 1–2. IEEE (2017)
41. Silvescu, A., Reinoso-Castillo, J., Honavar, V.: Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous biological data sources. In: In Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources. pp. 1–11 (2001)
42. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J., Wang, K.: An overview of microsoft academic service (mas) and applications. In: Proceedings of the 24th International Conference on World Wide Web. pp. 243–246 (2015)

43. Sivasubramaniam, A., Debnath, S., Li, H., Lee, W.C., Bolelli, L., Giles, C.L., Zhuang, Z., Councill, I.G.: Learning metadata from the evidence in an on-line citation matching scheme. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 276–285. IEEE (2006)
44. Soergel, D.: Digital libraries and knowledge organization. In: Semantic Digital Libraries, pp. 9–39. Springer (2009)
45. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH. pp. 484–496 (2020)
46. Tahamtan, I., Afshar, A.S., Ahamdzadeh, K.: Factors affecting number of citations: a comprehensive review of the literature. Scientometrics **107**(3), 1195–1225 (2016)
47. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1285–1293 (2012)
48. Traub, M.C., Van Ossenbruggen, J., Hardman, L.: Impact analysis of ocr quality on research tasks in digital archives. In: International Conference on Theory and Practice of Digital Libraries. pp. 252–263. Springer (2015)
49. Vahdati, S., Palma, G., Nath, R.J., Lange, C., Auer, S., Vidal, M.E.: Unveiling scholarly communities over knowledge graphs. In: International Conference on Theory and Practice of Digital Libraries. pp. 103–115. Springer (2018)
50. Wang, H., Tan, M., Yu, M., Chang, S., Wang, D., Xu, K., Guo, X., Potdar, S.: Extracting multiple-relations in one-pass with pre-trained transformers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1371–1377. ACL, Florence, Italy (Jul 2019)
51. Wang, W., Xia, F., Wu, J., Gong, Z., Tong, H., Davison, B.D.: Scholar2vec: Vector representation of scholars for lifetime collaborator prediction. ACM Transactions on Knowledge Discovery from Data (TKDD) **15**(3), 1–19 (2021)
52. Weigl, D.M., Kudeki, D.E., Cole, T.W., Downie, J.S., Jett, J., Page, K.R.: Combine or connect: Practical experiences querying library linked data. In: Proceedings of the Association for Information Science and Technology. vol. 56, pp. 296–305. Wiley Online Library (2019)
53. Yang, K., He, L., Dai, X., Huang, S., Chen, J.: Exploiting noisy data in distant supervision relation 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. 3216–3225 (2019)
54. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the ACL (volume 2: Short papers). pp. 207–212 (2016)
55. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 19–27 (2015)
56. Zuo, Z., Zhao, K.: Understanding and predicting future research impact at different career stages—a social network perspective. Journal of the Association for Information Science and Technology **72**(4), 454–472 (2021)