Incorporating the Rhetoric of Scientific Language into Sentence Embeddings using Phrase-guided Distant Supervision and Metric Learning

Kaito Sugimoto[1]
[1]The University of Tokyo
Tokyo, Japan
kaito_sugimoto@is.s.u-tokyo.ac.jp

Akiko Aizawa[2,1]
[2]National Institute of Informatics
Tokyo, Japan
aizawa@nii.ac.jp

Abstract

Communicative functions are an important rhetorical feature of scientific writing. Sentence embeddings that contain such features are highly valuable for the argumentative analysis of scientific documents, with applications in document alignment, recommendation, and academic writing assistance. Moreover, embeddings can provide a possible solution to the open-set problem, where models need to generalize to new communicative functions unseen at training time. However, existing sentence representation models are not suited for detecting functional similarity since they only consider lexical or semantic similarities. To remedy this, we propose a combined approach of distant supervision and metric learning to make a representation model more aware of the functional part of a sentence. We first leverage an existing academic phrase database to label sentences automatically with their functions. Then, we train an embedding model to capture similarities and dissimilarities from a rhetorical perspective. The experimental results demonstrate that the embeddings obtained from our model are more advantageous than existing models when retrieving functionally similar sentences. We also provide an extensive analysis of the performance differences between five metric learning objectives, revealing that traditional methods (e.g., softmax cross-entropy loss and triplet loss) outperform state-of-the-art techniques.[1]

1 Introduction

Scientific articles explain new ideas or discoveries and attempt to convince readers of their validity and importance. A key characteristic that distinguishes these articles from other texts is their specific rhetorical structures. The most well-known example is the main section of a paper, organized as Introduction, Methods, Results, and Discussion. Several attempts have also been made to identify argumentative roles within a section (Swales, 1990; Teufel et al., 1999; Lauscher et al., 2018). For example, a sentence in a paper beginning with "little attention has been paid to ..." shows the background of the research, or more specifically, the lack of previous research on that topic. In our work, we collectively refer to this rhetorical aspect of scientific writing as a *communicative function*, following Kanoksilapatham (2005).

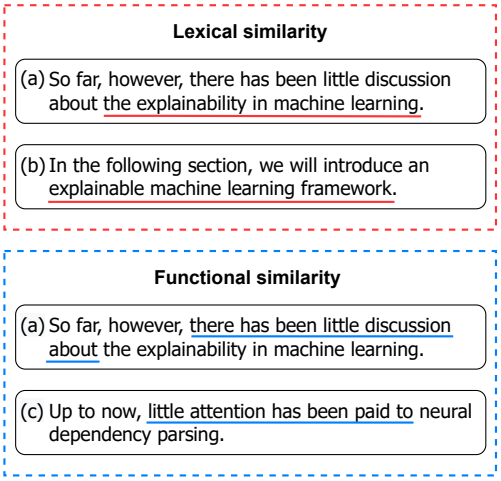

Figure 1: The upper panel shows an example of lexically similar sentences. Sentence (a) conveys the communicative function of "showing lack of previous work", whereas (b) conveys a different function, "showing the outline of the paper". In contrast, the lower panel shows a pair of functionally similar sentences.

Although previous studies have mainly focused on classifying sentences into a predefined set of communicative-function labels (Hirohata et al., 2008; Fisas et al., 2015; Cohan et al., 2019; Brack et al., 2022), we shift the focus to developing a sentence representation model for communicative functions. In other words, we consider sentence embeddings that can handle **functional similarity**, as opposed to lexical or semantic similarities

[1]Our code, data and trained models are publicly available at https://github.com/kaisugi/rhetorical_aspect_embeddings

Proceedings of the Third Workshop on Scholarly Document Processing, pages 54–68
October 12–17, 2022.

(Figure 1). There are two main reasons to prefer this approach: (i) The embedding model serves as an off-the-shelf tool to discover the most similar sentences to a query from a rhetorical perspective, which is beneficial for practical applications, including scientific document alignment (Zhou et al., 2020) and aspect-based scientific paper recommendation (Kobayashi et al., 2018; Chan et al., 2018). Such models can also contribute to writing assistance systems (Liu et al., 2016; Shioda et al., 2017) by suggesting sentences that have the same rhetorical feature as a query. (ii) Embeddings obtained from neural networks have shown the generalization ability to deal with the cases in which training and test sets do not share the same labels (i.e., open-set settings) (Musgrave et al., 2020; Geng et al., 2021). We argue that models for scholarly document processing (SDP) should perform well in open-set settings and generalize to unseen communicative functions, because there is no prepared list that covers all functional categories used in scientific articles.

In this paper, we introduce a new method for training a sentence representation model to capture functional similarity. We first address the scarcity of fine-grained datasets with communicative-function labels. Inspired by the success of distant supervision on low-resource natural language processing (Hedderich et al., 2021), we retrieve sentences from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020) and annotate labels based on simple text matching using an academic phrase dictionary, Academic Phrasebank (Davis and Morley, 2018). The resulting dataset, dubbed the **C**ommunicative-**F**unction-labeled **S**emantic **S**cholar **S**entence Dataset (CFS3), contains 100,016 sentences, classified into 77 function labels. We use this dataset to fine-tune SciBERT (Beltagy et al., 2019) with a metric learning loss so that functionally similar sentences come close together and dissimilar sentences are separated. As several recent studies (Musgrave et al., 2020; Boudiaf et al., 2020; Coria et al., 2020) have claimed that the performance of conventional metric learning losses (e.g., softmax cross-entropy loss) is comparable to or even better than that of state-of-the-art methods (e.g., ArcFace loss (Deng et al., 2019)), we also investigate whether these findings are valid in our settings.

We evaluate the trained model, named SciTORICSBERT[2], on sentence retrieval tasks designed to assess the rhetorical aspects of sentence representations. The experimental results show that our model is more suitable for retrieving functionally similar sentences than existing sentence representation models. We also observe that, in most cases, softmax cross-entropy loss yields better performance than other state-of-the-art methods. Furthermore, we train the same model using a limited number of communicative-function labels to better understand the generalizability of the trained models in open-set settings. The results reveal that the performance gain of conventional methods becomes even larger when the number of labels used for training becomes smaller.

Our contributions are as follows:

- We release CFS3, a distantly-labeled sentence dataset that includes 100K+ samples with 77 communicative-function labels.
- We present sentence embeddings that focus on the functional part of a sentence. Our model outperforms existing models in retrieving functionally similar sentences.
- We empirically demonstrate that the state-of-the-art metric learning methods do not improve performance on learning task-specific sentence embeddings.

2 Related Work

2.1 Argumentative Analysis of Scientific Texts

There is a large body of literature on assessing the argumentative status of scientific articles. Some notable schemes include move analysis (Swales, 1990) and argumentative zoning (Teufel et al., 1999). Another area of study is the annotation of communicative-function labels in abstracts using structured abstracts (Dernoncourt and Lee, 2017) or through crowdsourcing (Cohan et al., 2019; Huang et al., 2020).

Machine learning algorithms, such as conditional random fields (Hirohata et al., 2008), logistic regression, and support vector machines (Fisas et al., 2015), have been used to automatically classify sentences into function labels. Recently, SciBERT, a pre-trained language model on scientific texts, has pushed the limits of the classification accuracy (Cohan et al., 2019; Huang et al., 2020).

[2]The term SciTORICS was coined by Lauscher et al. (2018) to represent the rhetorical aspects of scientific writing.

2.2 Sentence Representation Models

Work on sentence embeddings can be divided into unsupervised and supervised methods. Conventional unsupervised models produce sentence embeddings by averaging each word or subword embedding from static or contextualized language models. This approach allows us to assess the lexical similarity of two sentences based on the distributional hypothesis (Harris, 1954). Recent supervised models trained on natural language inference (NLI) datasets (Conneau et al., 2017; Reimers and Gurevych, 2019; Gao et al., 2021) have shown significant improvements in semantic textual similarity (STS) tasks. These models can compute the semantic similarity of two sentences more faithfully than unsupervised models.

To the best of our knowledge, Iwatsuki et al. (2022) is the only study that investigated sentence representations for functional similarity. Their approach assigns different weights to word embeddings in the functional and non-functional parts of a sentence, whereas our proposed model eliminates the need to identify the functional part in advance.

2.3 Metric Learning

Metric learning (Kaya and Bilge, 2019) aims to learn a new mapping function from samples to vectors to reduce the distance between similar samples while increasing the distance between dissimilar samples. This training procedure is also called contrastive learning when the training data are annotated with pairwise labels (positive and negative pairs denote similar and dissimilar samples, respectively).

Triplet loss (or triplet margin loss) is one of the most studied learning methods. Neural networks associated with a triplet loss are known as "triplet networks", and they have been used in several applications, such as face recognition (Schroff et al., 2015), person re-identification (Hermans et al., 2017), sentence-level similarity learning (Ein Dor et al., 2018; Reimers and Gurevych, 2019), and document-level similarity learning (Cohan et al., 2020). Another classical approach is softmax cross-entropy loss. Although this loss is typically chosen for classification tasks, several studies have used it to train embedding models (Sun et al., 2014; Boudiaf et al., 2020).

Recently, much research has been devoted to designing loss functions to learn effective visual representations (Musgrave et al., 2020). These loss functions have been successfully applied to learning textual information, such as sentences (Yan et al., 2021; Giorgi et al., 2021; Kim et al., 2021; Gao et al., 2021), dialogues (Liu et al., 2021a), social media behaviors (Andrews and Bishop, 2019), and biomedical entities (Liu et al., 2021b). However, some studies have also shown that state-of-the-art loss functions do not necessarily outperform classical methods (Musgrave et al., 2020; Boudiaf et al., 2020; Coria et al., 2020).

3 Methods

Our approach can be roughly divided into two parts: phrase-guided distant supervision (Sections 3.1 and 3.2) and metric learning (Section 3.3), as illustrated in Figure 2.

3.1 Acquisition of Labeled N-gram List

Academic Phrasebank[3] is an online public database of generic academic phrases (Davis and Morley, 2018). Based on the observation that specific (formulaic) phrases serve as key markers for communicative functions (Swales, 1990), the database identifies 80 functions according to the main sections of a paper and samples approximately 20 phrases for each.

Our motivation is to utilize Academic Phrasebank for annotating sentences with communicative functions. Prior research has also leveraged this database to label sentences (Iwatsuki and Aizawa, 2021). However, their study relied on manual phrase extraction and annotation to maintain the quality of the labeling process. In contrast, we pursue a fully automated approach to create a larger, finer-grained dataset.

As the number of phrases in Academic Phrasebank is relatively small, we first perform data augmentation on the entire database using PPDB 2.0 (Pavlick et al., 2015) by randomly paraphrasing one noun, adjective, or adverb in a phrase. This results in a total of 30,505 phrases, which is approximately 20 times larger than the original.

The augmented phrases themselves are unsuitable for annotating sentences, because most of them are too lengthy to include specific content words that are irrelevant to communicative functions (e.g., "metabolism" in the phrase "X plays a vital role in the metabolism of ..."). We therefore extract every n-gram from the phrases. In

[3] https://www.phrasebank.manchester.ac.uk/

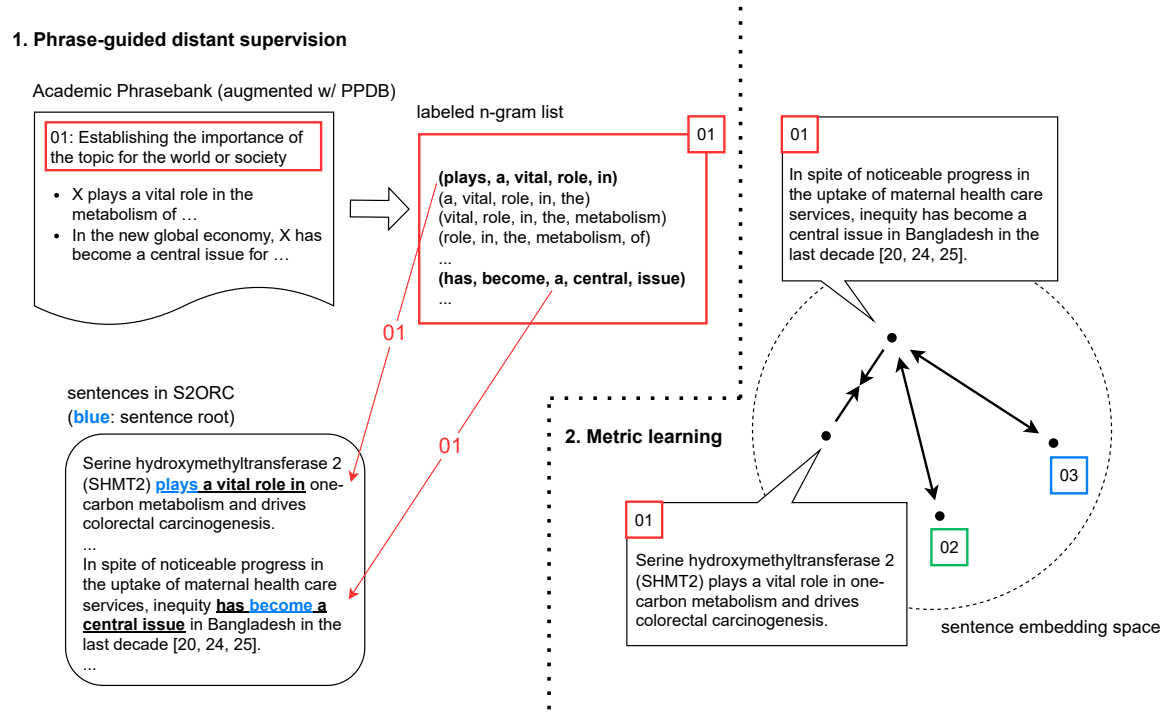

Figure 2: Overview of a combined approach of phrase-guided distant supervision and metric learning.

this study, we set $n = 5$.[4] We exclude from the list the lemmatized n-grams that have more than one label. As a result, we obtain 68,242 pairs of n-grams and their corresponding function labels. Although some of the n-grams (e.g., "vital role in the metabolism") still include content words, we find that they are negligible because such n-grams rarely retrieve sentences in Section 3.2.

3.2 Automatic Annotation of Sentences

We use the S2ORC (Lo et al., 2020) dataset to draw example sentences that contain specific n-grams. First, we randomly sample approximately 1M papers from S2ORC. Some are excluded during the preprocessing phase (see Appendix A for details). Then, we split each paper's abstract and body text into sentences using the NLTK tokenizer (Bird et al., 2009). This process produces approximately 19M sentences. Subsequently, for each labeled n-grams in Section 3.1, we inherit the same label for a sentence that satisfies the following constraints: (i) the sentence includes the n-gram, and (ii) the n-gram includes a root word in the dependency tree. The latter condition is derived from the observation that the functional part of a sentence often contains a sentence root (Iwatsuki et al., 2022). We use the spaCy (Honnibal et al., 2020) dependency parser to confirm whether the n-gram includes the root. This automatic annotation provides us with 100,016 labeled sentences.

Of the 80 function classes in Academic Phrasebank, three classes are assigned to no sentence; thus, the sentences are categorized into 77 classes. We name our dataset the Communicative-Function-labeled Semantic Scholar Sentence Dataset (CFS3). Table 1 contains randomly selected samples from CFS3. We find that the automatically-annotated sentences have expected function labels overall, except that, in the third sentence, the phrase "is interesting to note that" is not necessarily connected to the label "restating the result or one of several results", causing an annotation error.

3.3 Training with Metric Learning Loss

We train our embedding model using a metric learning framework to create a vector space in which sentences with similar functions have smaller distances, and those with different functions have longer ones. This trained model is referred to as SciTo^{RICS}BERT.

We begin from the pre-trained checkpoint of SciBERT (Beltagy et al., 2019) and take 768-dimensional embeddings from the `[CLS]` token

[4] We empirically determine that $n = 5$ is optimal. As we observe, for the case of $n < 5$, n-grams (e.g., "has been shown to") tend to be too generic to convey a specific communicative function. For the case of $n > 5$, on the other hand, n-grams often fail to retrieve any sentence.

in the last layer as the output. Then, we train the model with one of the five metric learning objectives mentioned below (the first two losses are conventional methods, while the rest are state-of-the-art methods that have been initially introduced in computer vision but also applied to natural language processing):

Softmax Cross-entropy Loss Let $\mathbf{x}_i \in \mathbb{R}^d$ denotes the output embeddings of the i-th sample, which belongs to the y_i-th communicative-function label ($1 \leq y_i \leq n$). Here, d is set to 768. We then minimize the following loss function:

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{W}_{y_i}^\top \mathbf{x}_i + b_{y_i})}{\sum_{j=1}^{n}\exp(\mathbf{W}_j^\top \mathbf{x}_i + b_j)}, \tag{1}$$

where $\mathbf{W}_j$ is the j-th column vector of the linear matrix, $\mathbf{W} \in \mathbb{R}^{d\times n}$, and b_j is the j-th element of the bias term, $\mathbf{b} \in \mathbb{R}^n$. N denotes the batch size.

Triplet Loss Triplets $\{a_i, p_i, n_i\}_{i=1}^N$ are collected from a training batch, provided that p_i has the same label as a_i, and that n_i has a different label.[5] We denote by $\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n \in \mathbb{R}^d$ the corresponding model outputs. Triplet loss is formulated as follows:

$$\mathcal{L}_2 = \frac{1}{K}\sum_{i=1}^{N}\max\left(\|\mathbf{x}_i^a - \mathbf{x}_i^p\|_2 - \|\mathbf{x}_i^a - \mathbf{x}_i^n\|_2 + m, 0\right), \tag{2}$$

where margin m denotes a hyperparameter, and K denotes the number of cases in which $\|\mathbf{x}_i^a - \mathbf{x}_i^p\|_2 - \|\mathbf{x}_i^a - \mathbf{x}_i^n\|_2 + m > 0$.

ArcFace Loss ArcFace loss (or additive angular margin loss) (Deng et al., 2019; Andrews and Bishop, 2019) modifies the softmax cross-entropy loss to make the learned embeddings more discriminative between classes.

We define $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{W}_j \in \mathbb{R}^d$ $(1 \leq j \leq n)$, which is similar to softmax cross-entropy loss. Let $\theta_j = \arccos\left(\frac{\mathbf{W}_j^\top \mathbf{x}_i}{\|\mathbf{W}_j\|_2\|\mathbf{x}_i\|_2}\right)$ be the angle between the output vector and the j-th column vector of the weight matrix. Then, ArcFace loss is defined as follows:

$$\mathcal{L}_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s\cos\theta'_{y_j})}{\sum_{j=1}^{n}\exp(s\cos\theta'_j)}, \quad \text{s.t. } \theta'_j = \begin{cases}\theta_j + m & (j = y_i)\\ \theta_j & (j \neq y_i)\end{cases}, \tag{3}$$

where angular margin m and scale s are hyperparameters.

MS Loss Multi-similarity (MS) loss (Wang et al., 2019; Liu et al., 2021b) considers multiple types of similarities for a pair, aiming to generalize previous loss functions.

Let $\mathbf{S} \in \mathbb{R}^{N\times N}$ be a similarity matrix whose

[5] In our work, all possible triplets are used for training without negative sampling.

Optical Flow: The estimation of optical flow is a classic problem in computer vision [18, 24] . **(02: Establishing the importance of the topic for the discipline)**
Initial and final nutrient concentrations, and significance between time points within each treatment group (t-test, p < 0.05) are shown in Figure 1 . **(48: Referring to data in a table or chart)**
It is interesting to note that one obtains Re J = 0 if cos ŏ3b1 e = 0 and U 0 is tri-bimaximal (t a = 1, **(63: Restating the result or one of several results)**
Method: A total of 104 participants (44 SZ patients and 60 age-and gender-matched healthy controls (HC)) were recruited for this study. **(36: Describing the characteristics of the sample)**
It has been suggested that dietary Zn is mostly absorbed in the duodenum, ileum, and jejunum by active transport through ZIP4 [48] . **(22: Previous research: what has been established or proposed)**
It has been suggested that bacteria may use hemolysin to obtain nutrients from the host cells (e.g., irons released from lysed red blood cells) [35] . **(22: Previous research: what has been established or proposed)**
This finding is consistent with other analyses, indicating that Tu-138 cells are more sensitive to E2F-1-induced apoptosis than are Tu-167 cells. **(65: Comparing the result: supporting previous findings)**
The modified QPM and the Delta method were used to analyse the data for each calendar month. **(47: Referring back to the research aims or procedures)**
It has been argued that the purposeful inclusion of social work values in social work research is one of its distinguishing features (Shaw et al., 2006) . **(22: Previous research: what has been established or proposed)**
Statistical analysis was performed using unpaired two-tailed Student's t-test where *P<0.05; **P<0.01. **(45: Describing the process: statistical procedures)**

Table 1: Ten randomly-selected examples from the CFS3 dataset. Function labels and corresponding n-grams are shown in bold and underlined, respectively.

Datasets		#sentences	#labels
CF-labeled	Introduction	773	11
	Methods	468	6
	Results	521	6
	Discussion	781	9
CSAbstruct		1,349	5
PubMed-RCT		30,135	5

Table 2: Dataset statistics.

(i, j)-th element satisfies $\mathbf{S}_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$, where $\mathbf{x}_i$ is the i-th model output in a N-sized training batch.

We regard a pair of two in-batch samples with the same label to be positive, and otherwise negative. We denote the sets of indices of positive and negative pairs by $\mathcal{P}$ and $\mathcal{N}$, respectively. The training objective is formulated as follows:

$$\mathcal{L}_4 = \frac{1}{N}\sum_{i=1}^{N}\Bigg\{\frac{1}{\alpha}\log\Big[1+\sum_{\substack{j=1,\\(i,j)\in\mathcal{P}}}^{N}\exp(-\alpha(\mathbf{S}_{ij}-\lambda))\Big] + \frac{1}{\beta}\log\Big[1+\sum_{\substack{j=1,\\(i,j)\in\mathcal{N}}}^{N}\exp(\beta(\mathbf{S}_{ij}-\lambda))\Big]\Bigg\}, \tag{4}$$

where α, β, and λ are hyperparameters.

NT-Xent Loss Normalized temperature-scaled cross-entropy (NT-Xent) loss (Chen et al., 2020; Giorgi et al., 2021) takes a form similar to softmax cross-entropy loss, but it differs in that it maximizes the similarity of a positive pair.

We define $\mathbf{S} \in \mathbb{R}^{N\times N}, \mathcal{P}, \mathcal{N}$ in the same manner as MS loss. NT-Xent loss can be expressed as follows:

$$\mathcal{L}_5 = -\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\log\frac{\exp(\mathbf{S}_{ij}/T)}{\exp(\mathbf{S}_{ij}/T)+\sum_{\substack{k=1,\\(i,k)\in\mathcal{N}}}^{N}\exp(\mathbf{S}_{ik}/T)}, \tag{5}$$

where temperature T is a hyperparameter.

4 Experiments

4.1 Settings

Task Description We conduct sentence retrieval tasks on communicative-function labeled datasets to see how successfully SciTo^{RICS}BERT contains rhetorical features. This task begins by converting all the sentences in a dataset into embeddings using a given representation model. We select one sentence as a query and regard the others as references. We then retrieve the nearest neighbors of the query and evaluate whether the extracted sentences have the same label as the query.[6] This procedure is repeated for the entire dataset, and the performance scores are averaged.

Evaluation Datasets We employ three datasets: **the CF-labeled sentence dataset** (Iwatsuki and Aizawa, 2021), **CSAbstruct** (Cohan et al., 2019), and **PubMed-RCT** (Dernoncourt and Lee, 2017). The CF-labeled sentence dataset is manually annotated with communicative-function labels for each section of papers from multiple disciplines. The other two datasets collect sentences from the abstracts of the computer science and biomedical domains, respectively. We report the dataset statistics in Table 2. Note that with CSAbstruct and PubMed-RCT, sentences in scientific abstracts are classified into one of the five categories {BACKGROUND, OBJECTIVE, METHOD, RESULT, CONCLUSION (OTHER)}, the granularity of which is much coarser than that of the CF-labeled sentence dataset (32 labels total).

Evaluation Metrics We use two evaluation metrics: precision at 1 (P@1) and mean average precision at R (MAP@R) (Musgrave et al., 2020). Whereas P@1 focuses on the top retrieval result, MAP@R measures the overall retrieval quality.[7] For SciTo^{RICS}BERT, we report the average results from five trained models with different random seeds.

Baselines We compare SciTo^{RICS}BERT with unsupervised language models, including average GloVe embeddings (Pennington et al., 2014), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). Other baselines include domain-specific language models, such as SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2020).[8] We also compare SciTo^{RICS}BERT with SRoBERTa (Reimers and Gurevych, 2019) and (supervised) SimCSE-RoBERTa (Gao et al., 2021), which are both fine-tuned RoBERTa models on NLI datasets.

Training Details To train SciTo^{RICS}BERT, we split our CFS3 dataset into a training and validation set at a ratio of 4:1. As the dataset is imbalanced,

[6] All the embeddings are L2 normalized beforehand.

[7] R denotes the total number of references with the same label as the query.

[8] Regarding transformer-based unsupervised models, we take the average of their last hidden layers.

Model	Introduction		Methods		Results		Discussion		Avg.	
	P@1	MAP	P@1	MAP	P@1	MAP	P@1	MAP	P@1	MAP
GloVe avg.	.391	.073	.462	.089	.484	.125	.325	.058	.415	.086
$\text{BERT}_{\text{base}}$ avg.	.523	.099	.434	.099	.511	.140	.361	.063	.457	.100
$\text{RoBERTa}_{\text{base}}$ avg.	.507	.106	.451	.102	.557	.152	.392	.068	.477	.107
SciBERT avg.	.604	.151	.526	.117	.612	.156	.483	.093	.556	.129
PubMedBERT avg.	.547	.134	.521	.108	.570	.140	.435	.078	.518	.115
$\text{SRoBERTa}_{\text{base}}$	.422	.099	.325	.075	.501	.148	.335	.072	.396	.099
$\text{SimCSE-RoBERTa}_{\text{base}}$	.551	.165	.429	.095	.511	.162	.403	.088	.474	.128
SciTo.BERT (Softmax loss)	.857	**.537**	.765	.375	**.866**	**.501**	**.741**	**.334**	**.807**	**.437**
SciTo.BERT (Triplet loss)	**.858**	.514	**.776**	.368	.855	.494	.734	.329	.806	.426
SciTo.BERT (ArcFace loss)	.840	.513	.741	.378	.845	.462	.708	.301	.783	.414
SciTo.BERT (MS loss)	.829	.485	.721	.354	.832	.475	.684	.314	.767	.407
SciTo.BERT (NT-Xent loss)	.839	.511	.741	**.385**	.838	.494	.708	.312	.781	.425

Table 3: Precision@1 and MAP@R scores for sentence retrieval tasks on the CF-labeled sentence dataset. The best-performing scores are highlighted in bold. The underlined scores are the highest among the baseline scores.

Model	CS		PubMed	
	P@1	MAP	P@1	MAP
GloVe avg.	.445	.124	.627	.167
BERT avg.	.553	.166	.681	.196
RoBERTa avg.	.523	.159	.681	.185
SciBERT avg.	.563	.169	.700	.204
PubMedBERT avg.	.553	.169	.694	.213
SRoBERTa	.480	.136	.566	.143
SimCSE-RoBERTa	.529	.164	.646	.187
SciTo.BERT				
(Softmax loss)	**.616**	**.226**	**.761**	**.325**
(Triplet loss)	.599	.214	.760	.324
(ArcFace loss)	.576	.205	.748	.307
(MS loss)	.583	.191	.739	.300
(NT-Xent loss)	.591	.216	.752	.324

Table 4: Precision@1 and MAP@R scores in sentence retrieval tasks on CSAbstruct and PubMed-RCT.

we follow stratified random sampling to ensure that both sets have similar label distributions. We measure the MAP@R score in the validation dataset for each epoch and select the best-performing model. The maximum number of epochs is set to five. See Appendix B for further detailed configurations.

4.2 Overall Results

We present the evaluation results for the CF-labeled sentence dataset in Table 3 and the other two datasets in Table 4.

Among the baseline models, SciBERT and PubMedBERT achieve the highest average scores. These domain-specific models consistently outperform BERT, indicating that pre-training on scientific texts provides *distributional functions* (i.e., words that occur in similar contexts have similar functions). As for supervised models, both SRoBERTa and SimCSE-RoBERTa perform poorly, sometimes even worse than RoBERTa. This suggests that semantic similarity does not help compare sentences from a rhetorical perspective.

Turning to our proposed method, we find that SciTo.BERT yields substantial improvements over the baselines in all the datasets. On the CF-labeled sentence dataset, the model achieves approximately 0.25 points gain in P@1 and 0.30 points gain in MAP@R over the best baseline. This result is not surprising because the labels in the CF-labeled sentence dataset are similar to those in our CFS3. More importantly, SciTo.BERT also outperforms on CSAbstruct and PubMed-RCT, although these datasets are generated from abstracts and are thus annotated with more coarse-grained function labels than CFS3.

Regarding the metric learning loss, there is no clear evidence that state-of-the-art methods are more competitive than conventional methods. Although triplet and NT-Xent losses achieve slightly better performance on some subsets of the CF-labeled sentence dataset, softmax cross-entropy loss outperforms all other methods in CSAbstruct and PubMed-RCT.

To illustrate the efficacy of our method, we compare the sentences retrieved by SciBERT and SciTo.BERT on the Introduction subset of the CF-labeled sentence dataset in Table 5. As the examples show, SciTo.BERT successfully suggests similar sentences based on the functional part of the query sentence. Additional examples are presented in Appendices C and D.

Query sentence			The main **question addressed in this paper concerns** whether it is possible to achieve a comparable or even better accuracy using just a small and non-redundant set of subtrees.
(Query function)			(Showing the outline of the paper)
SciBERT avg.		#1	The main challenge is the search problem, which is to find an optimal parse tree among all that can be constructed with any word choice and order from the set of input words.
	✓	#2	Another **issue addressed in this paper is** automatic construction of a lexicon for verbs related to activities and events.
		#3	Thus, the aim of this paper is to find an appropriate level of comparison for the combinatorial properties of music and language, ideally, in a way that is independent of controversies specific to one or the other field.
SciticsBERT (ours)	✓	#1	The third **issue addressed in this paper concerns** the nature of the category to be formed.
	✓	#2	The **problem addressed in this paper is** how to model and capture temporal contexts and how to enhance NED with this novel asset.
	✓	#3	Another **issue addressed in this paper is** automatic construction of a lexicon for verbs related to activities and events.
Query sentence			Second, **it remains unclear under which circumstances** higher inertia of positive emotions (PE) is maladaptive.
(Query function)			(Showing limitation or lack of past work)
SciBERT avg.		#1	However, the notion of automaticity has been challenged by subsequent studies.
		#2	Consequently, narrowing down which constructs are tied to ego depletion will help in solving the current controversy surrounding the effect.
	✓	#3	Currently, **little is known about how** auditory distraction impacts upon metacognitive regulation of memory responses as captured by the [CITATION] framework.
SciticsBERT (ours)	✓	#1	However, despite the success of NNLMs on large datasets ([CITATION], [CITATION], [CITATION]), **it remains unclear whether** their advantages transfer to scenarios with extremely limited amounts of data.
	✓	#2	**It remains unclear whether** similar enhancements in creativity can be observed if negatively arousing music is used.
	✓	#3	However, the molecular mechanism of NTP-induced cancer cell death **remains unclear**.

Table 5: Examples of top-3 sentences retrieved by SciBERT and SciticsBERT. ✓ stands for the same function label as the query. For ease of comparison, we show phrases that appear to accord with the function in bold.

4.3 Generalizability Analysis

We now investigate whether SciticsBERT generalizes across scientific documents or only memorizes specific phrasal patterns that accord with the communicative functions in our CFS3 dataset. We randomly sample 10, 20, or 40 of the 77 function labels in CFS3, train the model using only those data, and measure the average P@1 and MAP@R scores on the CF-labeled sentence dataset.[9] We hypothesize that models that have good generalizability can successfully retrieve similar sentences when trained on a portion of CFS3.

The results are shown in Figures 3 and 4. We see that all the models show strong performance over the best baseline, even if they are trained with only ten labels. This suggests that SciticsBERT can, to some extent, handle functional similarity in general. We also observe that P@1 scores keep higher values than MAP@R when training labels are reduced, indicating that the model uses clues to find the most similar sentence, which is easy to learn and generalizes well.

Notably, conventional softmax cross-entropy and triplet losses perform even better than the other methods when the number of training labels decreases. This contradicts our expectation as the other methods have achieved state-of-the-art results on the open-set image recognition tasks, where training and test sets do not share the same labels. One possible explanation is that the number of labels in our CFS3 is too small to train state-of-the-art methods effectively, considering that those

[9]To align the number of training samples, we vary the maximum training epoch in inverse proportion to the number of training labels. We report the average results from five trained models with different training labels and random seeds.

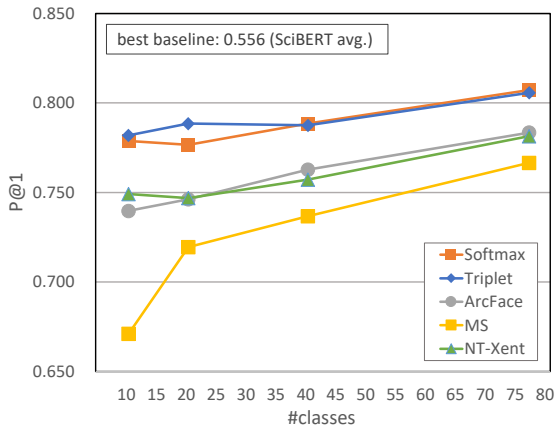

Figure 3: Effect of the number of classes on Precision@1 scores in the CF-labeled sentence dataset.

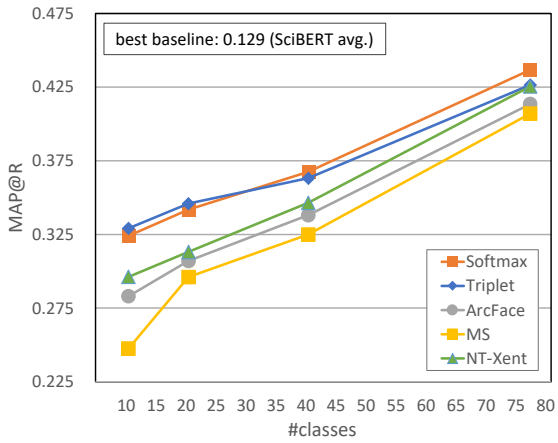

Figure 4: Effect of the number of classes on MAP@R scores in the CF-labeled sentence dataset.

methods are usually trained on a large-scale face recognition dataset containing thousands or millions of labels (e.g., the MS1MV2 dataset (Deng et al., 2019) contains 85K labels).

5 Conclusions and Future Work

This paper presents SciTo**RICS**BERT, a sentence representation model that recognizes the rhetorical aspects of scientific writing. The proposed model achieves more successful results than existing representation models in retrieving functionally similar sentences. We also provide empirical evidence that softmax cross-entropy loss is a strong baseline for learning task-specific sentence embeddings, which has practical implications for other studies on representation learning.

Future work should focus on improving our training methods using hard negatives (e.g., functionally dissimilar but lexically similar samples) and investigating our model in downstream applications.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. We are also grateful to Saku Sugawara for his insightful comments on an earlier version of the manuscript. This work was (partly) supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, Japan.

References

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. 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)*, pages 1684–1695.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. 2020. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 548–564.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, June 20 - 24, 2022*, pages 34:1–34:13.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A mixed initiative system for finding analogies between research papers. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):31:1–31:21.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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), pages 3693–3699.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Juan Manuel Coria, Sahar Ghannay, Sophie Rosset, and Hervé Bredin. 2020. A metric learning approach to misogyny categorization. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 89–94.

Mary Davis and John Morley. 2018. Facilitating learning about academic phraseology: teaching activities for student writers. *Journal of Learning Development in Higher Education*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. 2021. Recent advances in open set recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3614–3631.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Kenichi Iwatsuki and Akiko Aizawa. 2021. Communicative-function-based sentence classification for construction of an academic formulaic expression database. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3476–3497.

Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. 2022. Extraction and evaluation of formulaic expressions used in scholarly papers. *Expert Syst. Appl.*, 187:115840.

Budsaba Kanoksilapatham. 2005. Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24(3):269–292.

Mahmut Kaya and Hasan Sakir Bilge. 2019. Deep metric learning: A survey. *Symmetry*, 11(9):1066.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 243–251.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021a. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2396–2406.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021b. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yuanchao Liu, Xin Wang, Ming Liu, and Xiaolong Wang. 2016. Write-righter: An academic writing assistant system. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 4373–4374.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, volume 12370 of *Lecture Notes in Computer Science*, pages 681–699.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823.

Kent Shioda, Mamoru Komachi, Rue Ikeya, and Daichi Mochihashi. 2017. Suggesting sentences for ESL using kernel embeddings. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 64–68.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation from predicting 10, 000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1891–1898.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. 2020. Multilevel text alignment with cross-document attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5012–5025.

A Preprocessing in Dataset Construction

We conduct preprocessing before extracting texts from the S2ORC dataset. This phase proceeds in three steps. First, we exclude papers that lack venue or journal information in their metadata. Second, we exclude papers that do not contain body texts. Finally, we remove papers that are collected in one of the following corpora: *ACL anthology*, *Molecules*, *Oncotarget*, and *Frontiers in Psychology*. These four corpora are also used in the CF-labeled sentence dataset (Iwatsuki and Aizawa, 2021); thus, we consider that including them could cause data leakage. Note that the other two evaluation datasets contain papers in the computer science and biomedical domains, but we do not exclude them from the training data as some baselines such as SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2020) are already pre-trained on massive texts in those domains.

B Training Details

We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5. The batch size is set to 64. Following Hermans et al. (2017) and Musgrave et al. (2020), we adopt PK-style batches that first randomly sample P classes and then K instances for each class. We set $P = 64$ and $K = 1$ for Softmax and ArcFace losses and $P = 8$ and $K = 8$ for the others.

We conduct a hyperparameter search with fixed random seeds using the validation dataset, except for softmax cross-entropy loss. Table 6 lists hyperparameter configurations for each metric learning objective.

C Retrieval Examples by SciToricsBERT

Table 7 shows the retrieval examples by SciToricsBERT on the Introduction subset of the CF-labeled sentence dataset.

D A Case Study on Document Alignment

We showcase the utility of SciToricsBERT in the scenario of comparing different scientific papers. Specifically, we consider Devlin et al. (2019) and Lewis et al. (2020), which propose BERT and BART, respectively. We first retrieve texts from PDF files using S2ORC-doc2json (Lo et al., 2020), and split them into sentences using the NLTK tokenizer (Bird et al., 2009). Then, for each sentence

Loss	Hyperparameters
Triplet	$m \in \{0.025, 0.05^{\bullet}, 0.1, 0.2, 0.4\}$
ArcFace	$m \in \{0.1, 0.3, 0.5^{\bullet}\}, s \in \{16^{\bullet}, 32, 64\}$
MS	$\alpha \in \{1, 2^{\bullet}\}, \beta \in \{30, 40^{\bullet}, 50\}, \lambda \in \{0.5, 0.75^{\bullet}, 1.0\}$
NT-Xent	$T \in \{0.0125, 0.025, 0.05, 0.1^{\bullet}, 0.2\}$

Table 6: Values tested during the hyperparameter search. $^{\bullet}$ denotes those used for reporting the results.

in Lewis et al. (2020), we retrieve the most similar one from Devlin et al. (2019) using SCITORICS-BERT. We present a few selected examples in Table 8.

Query sentence			Dystrophin is an important protein for cytoskeletal structure and normal muscle function and plays a vital role in membrane stability and signaling [[CITATION]].
(Query function)			(Showing the importance of the topic)
SCITORICSBERT (ours)	✓	#1	VEGF is a major modulator of endothelial cell function, such as blood vessel formation during embryonic development, and plays a vital role in the proliferation, migration, and invasion of vascular endothelial cells [[CITATION]].
	✓	#2	Thrombin is an extracellular serine protease that plays a crucial role in the blood coagulation cascade, thrombosis, and hemostasis [[CITATION], [CITATION]].
	✓	#3	Copper is an essential element which plays a critical role in human metabolism.
Query sentence			From a computational standpoint, the main challenge is to ensure that the model scales well as the number of languages increases.
(Query function)			(Showing the main problem in the field)
SCITORICSBERT (ours)	✓	#1	, the main challenge is to detect the pattern without being distracted by background noise from other events.
	✓	#2	The main challenge is to maintain the continuity and coherence of the original text.
	✓	#3	The main challenge is to create a lexicon of dialect word forms and their associated probability maps.
Query sentence			Thus, in this paper we describe, for the first time, a straightforward synthesis of novel 1-(2'-α-O-D-glucopyranosyl ethyl) 2-arylbenzimidazoles via one-pot glycosylation of hydroxyethyl arylbenzimidazole aglycones and 2,3,4,6-tetra-O-benzyl 1-hydroxylglucose employing the Appel-Lee reagent [[CITATION], [CITATION]].
(Query function)			(Showing the importance of the research)
SCITORICSBERT (ours)		#1	The theoretical analysis developed in this paper aims to contribute to existing stage models of decision-making ([CITATION] [CITATION] [CITATION] [CITATION] [CITATION]).
		#2	Considering this, and in order to propose a greener route to fully epoxidized oligo-isosorbide glycidyl ethers, this paper reports a new protocol of heterogeneous ultrasound-assisted epoxidation in the presence of atomized sodium hydroxide.
	✓	#3	We argue for the first time that discourse parsing should be viewed as an extension of, and be performed in conjunction with, constituency parsing.
Query sentence			Recently, there has been a breakthrough in cancer immunotherapy against various cancer types by employing immune checkpoint blockade, particularly using antibodies directed against programmed death-ligand 1 (PD-L1) pathway members [[CITATION]].
(Query function)			(Showing brief introduction to the methodology)
SCITORICSBERT (ours)		#1	In recent years, there has been an increasing interest in controlled environment (CE) plant production which reduces variation related to climate, soil, and nutrition [[CITATION], [CITATION], [CITATION]], decreases contamination of samples by weeds, insects, and foreign matter [[CITATION]] and enhances the standardization of secondary metabolite production [[CITATION]].
		#2	In recent years, there has been an increasing interest in lichens as a potential source of pharmacologically bioactive agents for therapeutic treatments [[CITATION], [CITATION], [CITATION]].
	✓	#3	Non-human animal consciousness research has also witnessed groundbreaking advances in the study of contents of consciousness by employing perceptual rivalry paradigms and elucidating the effect of reversible thalamic and cortical inactivations.

Table 7: Examples of top-3 sentences retrieved by SCITORICSBERT. ✓ stands for the same function label as the query.

Query sentence from Lewis et al. (2020)	**Retrieved sentence from Devlin et al. (2019)**
We present BART, a denoising autoencoder for pretraining sequence-to-sequence models.	We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers.
The most successful approaches have been variants of masked language models, which are denoising autoencoders that are trained to reconstruct text where a random subset of the words has been masked out.	To pretrain word embedding vectors, left-to-right language modeling objectives have been used (Mnih and Hinton, 2009) , as well as objectives to discriminate correct from incorrect words in left and right context.
BART uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes (see Figure 1) .	Model Architecture BERT's model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al.
A key advantage of this setup is the noising flexibility; arbitrary transformations can be applied to the original text, including changing its length.	The advantage of these approaches is that few parameters need to be learned from scratch.
In total, BART contains roughly 10% more parameters than the equivalently sized BERT model.	By contrast, BERT BASE contains 110M parameters and BERT LARGE contains 340M parameters.
Unlike existing denoising autoencoders, which are tailored to specific noising schemes, BART allows us to apply any type of document corruption.	Unlike left-toright language model pre-training, the MLM objective enables the representation to fuse the left and the right context, which allows us to pretrain a deep bidirectional Transformer.
Because BART has an autoregressive decoder, it can be directly fine tuned for sequence generation tasks such as abstractive question answering and summarization.	As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial taskspecific architecture modifications.
Similar to BERT (Devlin et al., 2019), we use concatenated question and context as input to the encoder of BART, and additionally pass them to the decoder.	We use a gelu activation (Hendrycks and Gimpel, 2016) rather than the standard relu, following OpenAI GPT.
Following RoBERTa , we use a batch size of 8000, and train the model for 500000 steps.	We use a batch size of 32 and fine-tune for 3 epochs over the data for all GLUE tasks.
We mask 30% of tokens in each document, and permute all sentences.	In all of our experiments, we mask 15% of all WordPiece tokens in each sequence at random.
The most directly comparable baseline is RoBERTa, which was pre-trained with the same resources, but a different objective.	The most comparable existing pre-training method to BERT is OpenAI GPT, which trains a left-to-right Transformer LM on a large text corpus.
BART reduces the mismatch between pre-training and generation tasks, because the decoder is always trained on uncorrupted context.	BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953) .
Code and pre-trained models for BART are available at https://github.com/pytorch/fairseq and https://huggingface.co/transformers	The code and pre-trained models are available at https://github.com/google-research/bert.

Table 8: Example of document alignment using SCITORICSBERT.