Scientific Keyphrase Identification and Classification by Pre-Trained Language Models Intermediate Task Transfer Learning

Seo Yeon Park
Computer Science Department
University of Illinois at Chicago
spark313@uic.edu

Cornelia Caragea
Computer Science Department
University of Illinois at Chicago
cornelia@uic.edu

Abstract

Scientific keyphrase identification and classification is the task of detecting and classifying keyphrases from scholarly text with their types from a set of predefined classes. This task has a wide range of benefits, but it is still challenging in performance due to the lack of large amounts of labeled data required for training deep neural models. In order to overcome this challenge, we explore pre-trained language models BERT and SciBERT with intermediate task transfer learning, using 42 data-rich related intermediate-target task combinations. We reveal that intermediate task transfer learning on SciBERT induces a better starting point for target task fine-tuning compared with BERT and achieves competitive performance in scientific keyphrase identification and classification compared to both previous works and strong baselines. Interestingly, we observe that BERT with intermediate task transfer learning fails to improve the performance of scientific keyphrase identification and classification potentially due to significant catastrophic forgetting. This result highlights that scientific knowledge achieved during the pre-training of language models on large scientific collections plays an important role in the target tasks. We also observe that sequence tagging related intermediate tasks, especially syntactic structure learning tasks such as POS Tagging, tend to work best for scientific keyphrase identification and classification.

1 Introduction

Scientific **K**eyphrase **I**dentification and **C**lassification (SKIC) is the task of identifying and classifying scientific terms in research papers. An effective keyphrase identification and classification system can benefit a wide range of natural language processing and information retrieval tasks including question answering (Quarteroni and Manandhar, 2006), question generation (Subramanian et al., 2018), and expert finding (Chen et al., 2015; Augenstein et al., 2017). Several works formulate SKIC as a classification task using neural methods and word embeddings (Luan et al., 2018; Augenstein et al., 2017; Liu et al., 2017) and show promising results compared to previous hand-crafted feature processing with sequence labeling such as Conditional Random Fields (Lee et al., 2017).

Scientific keyphrases are very diverse and this diversity leads to a burden for the creation of large dataset collections, which results in data scarcity problems for deep neural networks since domain experts are required to obtain reliable annotations for keyphrase identification and classification. In order to overcome the small size data problem of SKIC, Augenstein and Søgaard (2017) propose deep multi-task learning and reveal that several related tasks such as noun and verb phrase chunking, super-sense tagging, and multi-word expression identification are helpful to SKIC. However, the results of this approach are still low (Augenstein et al., 2017), which implies that there is room for improvement.

Recent works in natural language processing show that employing unsupervised pre-trained language models such as BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019), a variant of BERT that is pre-trained on scholarly texts, can bring substantial improvement in the performance of various Natural Language Understanding (NLU) tasks (Jiang and de Marneffe, 2019; Klein and Nabi, 2019). Furthermore, Phang et al. (2018) and Pruksachatkun et al. (2020) propose to further improve these language

Proceedings of the 28th International Conference on Computational Linguistics, pages 5409–5419
Barcelona, Spain (Online), December 8-13, 2020

models by using intermediate task transfer learning. Intermediate task transfer learning is a simple strategy of fine-tuning on a data-rich intermediate task before fine-tuning on the downstream target tasks. To this end, we investigate the performance of intermediate task transfer learning for SKIC. Specifically, we run our experiments on seven intermediate tasks on BERT and SciBERT and six target tasks (three for keyphrase boundary identification and three for keyphrase type classification) using three datasets of scientific papers that cover different domains, e.g., Materials Science, Physics, Computer Science, Natural Language Processing, and Artificial Intelligence.

Our contributions are summarized as follows: First, we build a transfer learning framework employing a diverse range of intermediate tasks covering sequence tagging with semantic and syntactic aspects, and natural language inference. We achieve competitive performance over both strong baselines and previous works. While transfer learning using SciBERT successfully achieves improvement in performance, we observe that intermediate transfer learning using BERT causes catastrophic forgetting (Chronopoulou et al., 2019) with a significant performance deterioration. Second, we empirically observe that specific intermediate *task types* are more useful as intermediate tasks for SKIC. Specifically, sequence tagging tasks such as POS tagging are preferable than natural language inference tasks such as entailment recognition. Third, we provide a qualitative analysis of extracted keyphrases and show that our transfer learning successfully returns keyphrases. This qualitative analysis proves the reliability of our proposed methods.

2 Related Work

Scientific keyphrase identification and classification (SKIC) was proposed at SemEval 2017 Task 10 as *Task A*: Keyphrase Identification, and *Task B*: Keyphrase Classification (Augenstein et al., 2017). The authors of this shared task proposed to employ the BIO schema, which refers to the beginning (B), inside (I), or outside (O) of keyphrases, respectively. Most of the earlier approaches to SKIC rely on hand-crafted linguistic features. For example, Lee et al. (2017) and Marsi et al. (2017) employ syntactic and semantic features such as Part-of-Speech tags and word lemmas, as input to Conditional Random Fields (CRFs). Liu et al. (2017) formulate SKIC as a supervised multi-class classification problem. They exploit pre-trained word embeddings and linguistically inspired features, e.g., noun phrase features and orthographic features, which represent character and symbolic features of given tokens, as input to Support Vector Machines. With the success of neural models, recent works try to address SKIC using neural architectures while exploiting the BIO schema. Although both tasks, keyphrase identification and keyphrase classification according to their types, are very important, many works focused only on keyphrase extraction/generation or identification/segmentation (Meng et al., 2017; Xiong et al., 2019; Patel and Caragea, 2019; Alzaidy et al., 2019; Chen et al., 2020). The classification task is less explored possibly due to a lack of a large number of gold-label keyphrase classification datasets. Precisely, there are only a few publicly available datasets for keyphrase classification (QasemiZadeh and Schumann, 2016; Augenstein et al., 2017; Luan et al., 2018) and these datasets are small in size. To overcome the small dataset size problem, Augenstein and Søgaard (2017) proposed the multitask learning of SKIC. According to this work, they reveal that sequence tagging auxiliary tasks such as Chunking, Super-sense Tagging, Multi-words Expressions Identification, and FrameNet Target Identification are beneficial to SKIC within a multitask learning framework (with one auxiliary task at a time).

The recent unsupervised pre-trained language models such as BERT (Devlin et al., 2019) achieve state-of-the-art performance in many downstream NLP tasks, such as named entity recognition e.g., 95.5% F1 on the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). Han and Eisenstein (2019) apply BERT and use fine-tuning of BERT to reduce the vocabulary gap between canonical domains and historical domains, within an unsupervised approach. Moreover, the domain-specific BERT models such as SciBERT (Beltagy et al., 2019), produce better results compared to a general domain-based BERT with respect to scientific knowledge required tasks such as scientific term classification e.g., 64.57% F1 on the SciIE dataset (Luan et al., 2018). To exploit these pre-trained language models, Phang et al. (2018) propose a data-rich intermediate transfer learning and prove that this method provides a better starting point of target tasks. Extended from this, in our work, we propose the transfer of the intermediate tasks fine-tuning to overcome the scarcity of gold-labeled large SKIC datasets.

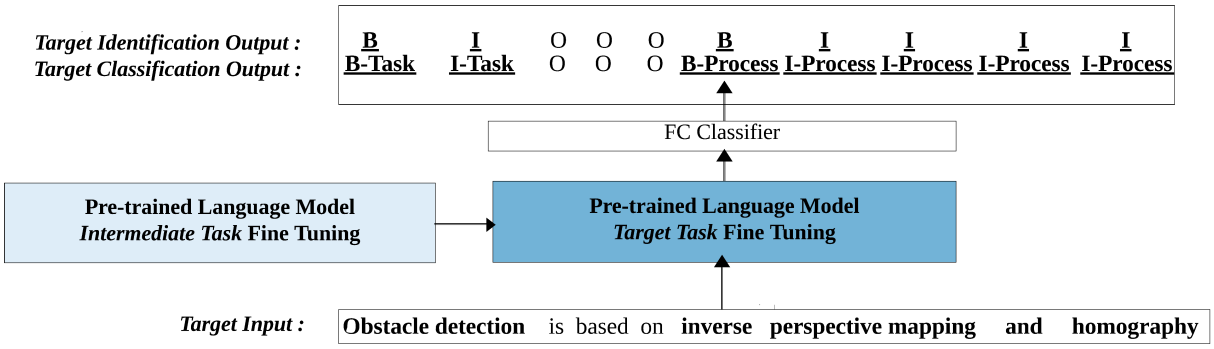

Figure 1: The overview of intermediate task transfer learning for keyphrase identification and classification by using pre-trained language models. The words from the input shown in bold represent keyphrases and their types are shown in the classification output.

3 Methodology

Given a document $d = \{w_1, w_2, \dots w_n\}$, where n denotes the length of the document and w_k is a word in d ($k = 1, \cdots, n$), the objective of SKIC is to identify and classify keyphrases of d by employing an output sequence $\{y_1, y_2, \dots y_n\}$, where y_k follows BIO scheme for pre-defined classes, e.g., task, material, process. In the BIO scheme, 'B' denotes the beginning word of a keyphrase, 'I' refers to the inside word of a keyphrase, and 'O' indicates all other words that are not part of a keyphrase. An example of input-output pairs of keyphrase identification and classification is shown in Figure 1.

3.1 Experimental Pipeline

Our transfer learning consists of two phases as shown in Figure 1: (1) Pre-trained language models BERT and SciBERT are fine-tuned respectively on a single intermediate task; (2) The pre-trained language models' fine-tuned-parameters are transferred to the target task and fine-tuned further on each target task. In target task fine-tuning, we apply different learning rates to retain knowledge that is acquired from intermediate tasks. Specifically, as shown in Figure 1, the target task's fully connected classifier uses a higher learning rate, while pre-trained language models' parameters are assigned as a lower learning rate to retain knowledge.

3.2 Intermediate Tasks

To explore what intermediate tasks are effective for keyphrase identification and classification, we choose various tasks. Intermediate tasks statistics are shown in Table 1. Note that for intermediate tasks that are related to sequence tagging, we report phrase-level instance numbers.

We choose seven intermediate tasks based on the following considerations: 1) Augenstein and Søgaard (2017) propose several auxiliary tasks where they improve performance of SKIC using a multitask learning framework; 2) Pruksachatkun et al. (2020) prove the usefulness of Natural Language Inferences (NLIs) in intermediate task transfer learning for general Natural Language Understanding. In our setting, document understanding is essential to identifying the precise keyphrases (or the important parts) for a document. In experiments, we use two NLI tasks where each task has different vocabulary sets, i.e., general domain (MNLI) and scientific domain (SciTail); 3) A traditional NLP sequence tagging task, i.e., Part-Of-Speech Tagging, to understand whether a model that acquire better ability of understanding grammatical structures improves the performance of SKIC. The detailed information of each intermediate task follows.

Supersense Supersense tagging (Johannsen et al., 2014) is the task of mapping semantic units such as compounds, phrase and several other linguistic terms to Wordnet's lexicographers classes using Semcor

Task Name	|Train|	|Dev|	|Class|	Task Type	Domain
Supersense	8,006	1,186	3	Sequence tagging	Newswire, Novel
MWEs	11,890	1,146	57	Sequence tagging	Online review
FrameNet	9,334	4,000	92	Sequence tagging	Open domain
Chunking	104,054	11,588	12	Sequence tagging	Wall Street Journal
SciTail	23,596	1,304	2	Natural language inference	Science exams
MNLI	392,703	20,000	3	Natural language inference	Fiction, Letters, Telephone Speech
POS Tagging	334,180	54,138	15	Sequence tagging	Wikipedia, Talks, Literature

Table 1: The overview of intermediate tasks in our experiments.

3.0 corpus.[1] In this work, we focus on noun supersense classes which are *Person, Location* and *Group*. For example, in a given sentence *'Boston Red Sox outfielder Jackie Jensen said he played baseball on monday night'*, the supersense of *Boston Red Sox* is *Group*, and that of *Jackie Jensen* is *Person*.

MWEs Multiword Expressions Identification and Classification (Schneider and Smith, 2015) is the task of detecting and classifying idiosyncratic noun and verb combinations to literal categories, which are high-level ontological semantic classes, using Streusle corpus.[2] For example, in a given sentence *'We have been blessed to find elite flyers online and would not use anyone else to handle our postcards posters etc.'*, examples of high-level ontological semantic classes are: for *blessed* is *verb.cognition*, for *find* is *verb.cognition*, for *use* is *verb.social* and for *postcards posters* is *noun.artifact*.

FrameNet FrameNet Target Identification (Das et al., 2014) is the task of detecting semantic frame evoking word sequences or words in the FrameNet 1.7 corpus, which contains lexical and predicate-arguments. Frames are words or phrases that describe events, relations, objects and the participants in it. For example, a target such as *moist* in a sentence evokes the frame *Being* as state.

SciTail SciTail (Khot et al., 2018) is the task of recognizing the entailment of a hypothesis that is constructed from a science question and its corresponding answer by employing the premise. The dataset is collected by crowd-sourcing and by multiple-choice science questions from 4th-grade and 8th-grade exams. For example, the relation between the following two sentences, *'Neurons receive information from dendrites which are then passed to the soma cell body.'* and *'Dendrites from the cell body receives impulses from other neurons.'* is labeled as *Entail*.

MNLI Multi-Genre Natural Language Inference (Williams et al., 2018) is the task of determining textual entailment in sentence pairs across a variety of genres of written and spoken English, ranging from fiction to face-to-face conversations. For example, the relation between the following two sentences, '*This approach provides perhaps a better technique for isolating the actual costs of the emissions caps.*', '*There is no way to estimate the actual cost of emissions caps.*' is labeled as *Contradiction*.

Chunking Text Chunking is the task of detecting the chunks of words in CoNLL-2000 shared dataset (Tjong Kim Sang and Buchholz, 2000). The chunk tags include various grammatical classes such as noun phrase, verb phrase, prepositional phrase and these tags follow the BIO schema. For a sentence *'He reckons the current account deficit will narrow to only #1.8 billion in September'* can be annotated as,*'[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]'*, where NP refers to noun phrase, VP refers to verb phrase, PP refers to prepositional phrase.

POS Tagging POS Tagging is the task of labeling each word in a sentence with its part of speech tag. In this work, we employ the English POS tagging annotation collection of universal dependency parsing (McDonald et al., 2013). As an example, the following sentence, *'Aesthetic appreciation and spanish art.'*, has the grammar class sequence as *['ADJ', 'NOUN', 'CCONJ', 'ADJ', 'NOUN']*.

[1] https://web.eecs.umich.edu/ mihalcea/downloads.html

[2] https://github.com/nert-nlp/streusle

Task Name	|Train|	|Dev|	|Test|	|Class|	|Avg KP per Doc|	|Avg Doc Word Count|	Domain
SemEval 2017	5,992	1,076	1,817	3	18	162	Computer Science, Physics, and Material Science
ACL RD-TEC 2.0	1,930	214	1,088	7	12	107	Natural Language Processing
SciIE	5,712	641	1,677	6	17	120	Artificial Intelligence

Table 2: The statistics of target tasks in our experiments.

3.3 Target Tasks

Table 2 shows the characteristics of our target tasks. The detailed information is presented below.

SemEval 2017 Task 10[3] SemEval 2017 (Augenstein et al., 2017) is a scientific keyphrase boundary identification and classification dataset. The SemEval 2017 dataset covers three domains: materials science, physics, and computer science. The dataset has three pre-defined classes which are *Process*, *Task*, and *Material*.

ACL RD-TEC 2.0[4] ACL (QasemiZadeh and Schumann, 2016) is a term and entity categorization task in scientific text. The ACL dataset covers the domain related to natural language processing. The ACL dataset has seven pre-defined classes, which are *Technology (Tech), Tool, Language Resources (Lr), Language Resources Product (Lr-prod), Measurement, Model*, and *Other*. Note that the class *Other* here is different from the class *Other* in the BIO scheme that represents non-keyphrase.

SciIE[5] SciIE (Luan et al., 2018) covers the task of detecting scientific entities, their relations, and coreference clusters. In this work, we focus on term identification and classification. The SciIE dataset covers the domain related to artificial intelligence in computer science. The SciIE dataset has six pre-defined classes, which are *Generic, Metric, Method, Task, Material*, and *OtherScientificTerm (Other)*.

4 Experiments

Baseline We compare the intermediate task transfer learning with the following baselines.

- **BiLSTM (Augenstein and Søgaard, 2017)** A 3-layer BiLSTM with SENNA embeddings[6] for each target task.
- **BiLSTM-MTL (Augenstein and Søgaard, 2017)** Multitask learning of 3-layer BiLSTMs with SENNA embeddings.
- **BERT (Devlin et al., 2019)** BERT$_{base}$ fine-tuning on a single target task.
- **SciBERT (Beltagy et al., 2019)** SciBERT fine-tuning on a single target task.

Experimental Setup For the SemEval 2017 and SciIE, we use the published train, validation, and test sets. For the ACL RD-TEC 2.0 dataset, we perform 60/10/30 split to create the train, validation, and test sets. We estimate models' hyper-parameters via a grid search over combinations. The ranges of parameters we explore are the following: Batch size [1,4,8,16,32,64], learning rate [0.01, 0.0001], momentum [0.5, 0.9]. The information about the best hyper-parameters setting is as follows. For each target task, we use the batch size 1. The intermediate tasks of Supersense, MWEs, FrameNet use batch size 4, while Chunking uses 32, POS Tagging uses 64, SciTail uses 16, and MNLI uses 4. For BERT and SciBERT fine-tuning, we use SGD optimizer with learning rate 5e-3 with momentum 0.9. In intermediate-task transfer learning, we train intermediate tasks with learning rate 5e-3 without applying momentum. We set a higher learning rate as 5e-3 and set a lower learning rate as 5e-4, both with momentum 0.9. We run

[3]https://scienceie.github.io/
[4]https://github.com/languagerecipes/acl-rd-tec-2.0
[5]http://nlp.cs.washington.edu/sciIE/
[6]https://ronan.collobert.com/senna/

	SemEval 2017 Task 10		ACL RD-TEC 2.0		SciIE	
	Identification	**Classification**	**Identification**	**Classification**	**Identification**	**Classification**
BiLSTM (Augenstein and Søgaard, 2017)	67.71	38.01	81.85	58.51	72.33	58.05
BiLSTM-MTL w/ Supersense	63.93	43.54	81.36	58.95	72.65	54.33
BiLSTM-MTL w/ MWEs	72.42	45.49	80.69	56.87	72.92	55.21
BiLSTM-MTL w/ FrameNet	65.18	45.24	81.68	58.89	70.44	52.60
BiLSTM-MTL w/Chunking	63.96	42.86	81.37	57.84	75.40	59.43
BERT (Devlin et al., 2019)	60.40	46.82	79.50	51.67	81.02	65.44
Transfer From Supersense/ BERT	63.52	49.02	71.03	44.78	75.53	57.40
Transfer From MWEs/ BERT	59.22	45.92	70.70	43.13	75.22	57.05
Transfer From FrameNet/ BERT	61.81	46.06	71.04	42.64	73.82	55.06
Transfer From SciTail/ BERT	60.11	47.81	63.77	39.87	74.59	55.73
Transfer From MNLI/ BERT	49.44	44.52	64.51	40.22	73.21	54.42
Transfer From Chunking/ BERT	65.87	47.90	77.72	45.54	75.81	55.18
Transfer From POS Tagging/ BERT	59.64	42.08	76.97	45.77	75.87	52.89
SciBERT (Beltagy et al., 2019)	66.70	48.21	79.77	65.11	81.02	67.44
Transfer From Supersense/ SciBERT	70.39	51.58	81.62	67.75	**82.65**	70.20
Transfer From MWEs/ SciBERT	72.09	50.90	83.65	67.55	80.55	67.94
Transfer From FrameNet/ SciBERT	**73.89**	53.70	82.95	66.73	80.94	69.63
Transfer From SciTail/ SciBERT	68.49	55.52	82.89	68.20	81.28	67.57
Transfer From MNLI/ SciBERT	70.35	54.08	76.51	63.16	76.51	**75.23**
Transfer From Chunking/ SciBERT	72.07	53.75	83.28	66.35	77.08	66.20
Transfer From POS Tagging/ SciBERT	70.90	**56.90**	**88.01**	**69.90**	78.62	66.64

Table 3: The results of BERT and SciBERT intermediate task transfer learning in comparison with previous work. We use the micro-average F1 score. The results are highlighted with green (↑) and red (↓) with respect to BERT (middle) and SciBERT (bottom). Underlined scores are best within each group and bold scores are best overall.

training for a maximum of 10 epochs with negative log-likelihood loss. Aside from these details, we follow the SciBERT paper for all other training hyper-parameters. All experiments are done in p3.2xlarge settings of Amazon Web Services. All model parameters are estimated on the validation set of each task. We evaluate the performance of each model using phrase-level micro-averaged F1 and use the exact match metric (Kim et al., 2010).

5 Results and Analysis

5.1 Discussion

Table 3 presents our identification and classification results. We make the following observations.

First, we observe that SciBERT generally shows higher performance across all of our target tasks in comparison to BERT. Interestingly, BERT and SciBERT have different vocabulary set which only overlapped 42% (Beltagy et al., 2019). We posit that scientific knowledge that is available in SciBERT boosts the performance of SKIC. However, the fine-tuning of SciBERT on keyphrase identification for both SemEval 2017 and ACL datasets still remains challenging since the performance is lower than the best performance of the previous work (Augenstein and Søgaard, 2017). For example, the best baseline F1-score on SemEval 2017 keyphrase identification is 72.42%, whereas that of SciBERT fine-tuned on SemEval 2017 keyphrase identification achieves only 66.70%.

Second, our SciBERT transfer learning achieves the best results in all identification and classification target tasks. This implies that intermediate task fine-tuning leads to better starting points for the target task fine-tuning by injecting the knowledge related to SKIC. Specifically, in keyphrase identification and classification of the SemEval 2017 dataset, the model gains significant effectiveness when FrameNet is employed as an intermediate task (+7.19 F1), and when POS Tagging (+8.69 F1) is used as an intermediate task, respectively. For the ACL dataset, we observe that exploiting POS Tagging as an intermediate task achieves the best performance of both identification (+6.16 F1) and classification (+4.79 F1). This implies that a better syntactic understanding ability of the model induces performance improvement. The SciIE dataset performs the best with transfer from Supersense in keyphrase identification (+1.63 F1), and MNLI in keyphrase classification (+7.79 F1).

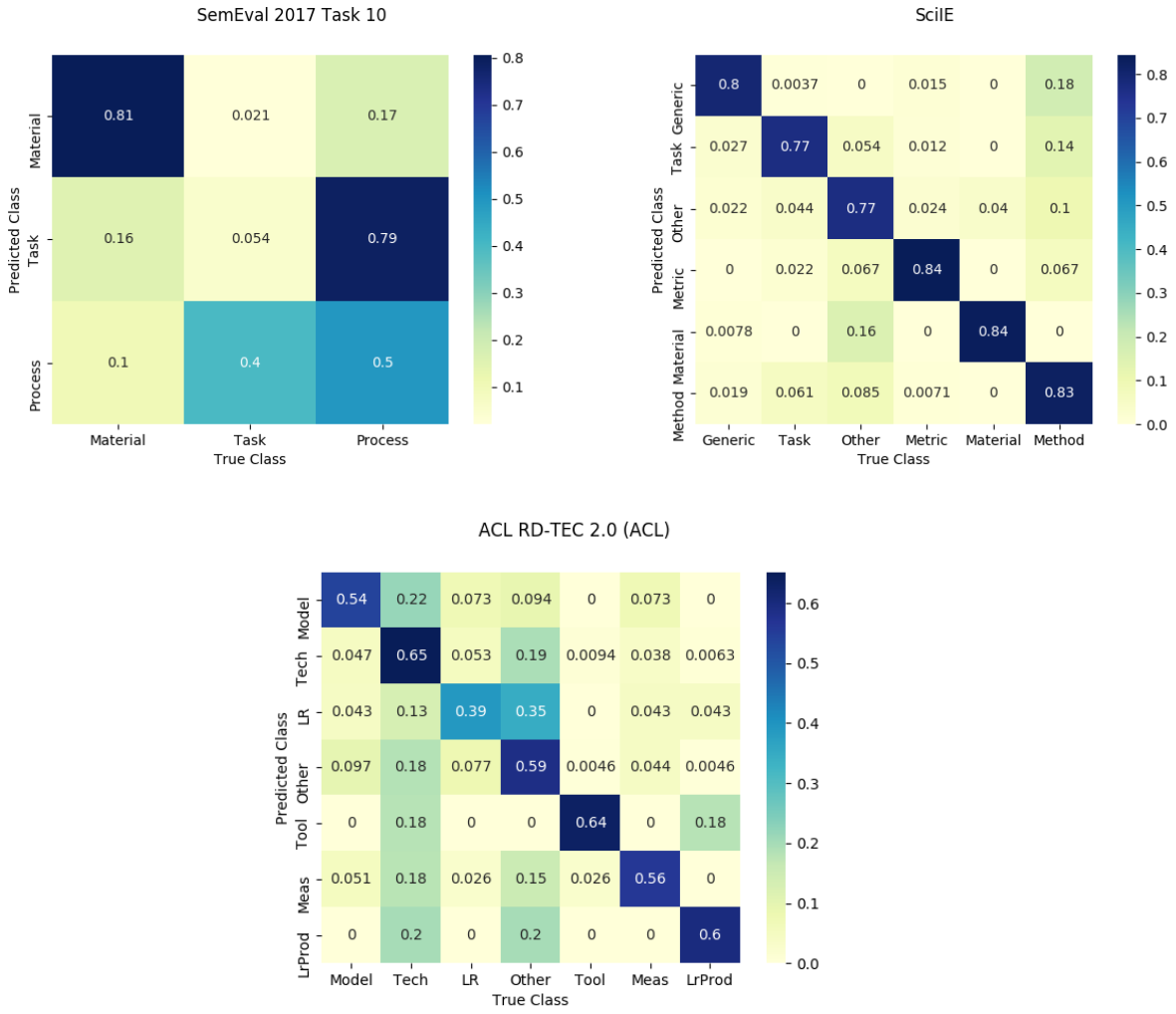

Figure 2: The keyphrase classification results confusion matrix visualization of each target dataset. The x-axis refers to true classes and the y-axis refers to predicted classes in each target task. The numbers are normalized by the count of predicted keyphrases.

More interestingly, the contribution of intermediate tasks to target tasks is different by their *task types*. When natural language inference tasks are employed as intermediate tasks, we generally observe performance degradation, while most of the sequence tagging tasks are consistently helpful for SKIC. Syntactically sequence tagging tasks such as chunking and POS Tagging are also generally helpful to SKIC. Another important observation is that the domain difference in natural language inference tasks (general and scientific domain) results in different outputs. For example, in all three datasets, transferring from MNLI in SciBERT shows better performance than transferring from SciTail in BERT. MNLI and SciTail lie on similar inference recognition settings but have different domains, which are general domains and scientific domains, respectively.

Third, we observe that BERT potentially suffers from catastrophic forgetting as opposed to SciBERT. In BERT transfer learning, most of the intermediate tasks fail to provide useful knowledge for the target tasks (i.e., they appear to make the fine-tuned models forget the good information learned within the pre-trained BERT), and hence, result in a severe deterioration of performance. For example, on the ACL RD-TEC 2.0 keyphrase identification task, when SciTail is employed as an intermediate task on BERT transfer learning, the performance is decreased by as much as -15.73 in F1, compared to single BERT fine-tuning. As we can see from Table 3, all cases of BERT intermediate task transfer learning in ACL and SciIE dataset degrade performance in comparison to BERT, while most of the cases in SciBERT transfer learning improve performance compared to that of SciBERT.

Article from SemEval 2017 Task 10
PV cells are one of the most promising technologies for conversion of incident solar radiation into electric power. However, this technology is still far from being able to compete with fossil fuel-based energy conversion technologies because of its relatively low efficiency and energy density. Theoretically, there are three unavoidable losses that limit the solar conversion efficiency of a device with a single absorption threshold or band gap Eg: (1) incomplete absorption, where photons with energies below Eg are not absorbed; (2) thermalization or carrier cooling, where solar photons with sufficient energy generate electron-hole pairs and then immediately lose almost all energy in excess of Eg in the form of heat; and (3) radiative recombination, where a small fraction of the excited states radioactively recombine with the ground state at the maximum power output (Hanna & Nozik, 2006; Henry, 1980). Taking an air mass of 1.5 as an example, for different band gap Eg these three losses can be calculated and the results are indicated by areas S1, S2, and S3 in Fig. 1. Note that the area under the outer curve is the solar power per unit area, and that only S4 can be delivered to the load.
Gold Keyphrases
('PV cells', 'Task') ('conversion of incident solar radiation into electric power', 'Process') ('incident solar radiation', 'Material') ('electric power', 'Material') ('fossil fuel-based energy conversion', 'Process') ('absorption', 'Process') ('incomplete absorption', 'Process') ('thermalization', 'Process') ('carrier cooling', 'Process') ('photons', 'Material') ('solar photons', 'Material') ('electron-hole pairs', 'Material') ('radiative recombination', 'Process') ('solar power', 'Process')
Keyphrase Outputs of Transfer from POS Tagging/ SciBERT
('PV cells', 'Material') ('conversion of incident solar radiation into electric power', 'Task') ('fossil fuel based energy conversion technologies', 'Material') ('absorption', 'Process') ('incomplete absorption', 'Process') ('carrier cooling', 'Process'), ('photons', 'Material') ('photons', 'Material') ('electron-hole pairs', 'Material') ('heat', 'Process') ('radiative recombination', 'Process') ('air mass', 'Material') ('solar', 'Material')

Table 4: The comparison between gold keyphrase and classified keyphrase outputs of SemEval 2017 Task 10 dataset.

5.2 Error Analysis

We visualize confusion matrices of the best performing SciBERT transfer learning keyphrase classification results in Figure 2. Specifically, for the SemEval 2017 Task 10 dataset and the ACL dataset, we plot the result of SciBERT transfer learning from POS tagging. For the SciIE dataset, we plot the result of SciBERT transfer learning from MNLI. The numbers in Figure 2 represent how many classified keyphrases belong to each true class. For example, in SemEval 2017 Task 10 dataset confusion matrix, the cell corresponding to row *Process* and column *Task* refers to the ratio of keyphrases predicted as *Task* but which should be classified as *Process* to the total number of keyphrases that are classified as *Task*. Consequently, each row in every confusion matrix sums up to 1.

We observe that SemEval 2017 data suffers from mis-classifying the class *Task* possibly due to the imbalanced data distribution. The distribution of categories in SemEval 2017 Task 10, *Process*, *Material*, *Task* is 50%/35%/15%, respectively. Moreover, our SciBERT transfer learning makes erroneous predictions between *Process* and *Task* categories due to the subjectivity of two classes as shown in Table 4. Table 4 presents an example article of the SemEval 2017 and its keyphrase classification comparison between gold labels and SciBERT transfer learning from POS Tagging. We observe that the keyphrase *conversion of incident solar radiation into electric power* is annotated as *Process*, while it is reasonable to think of it as *Task*, which is the output of our transfer learning model prediction. This type of error is not necessarily a shortcoming of SciBERT, but rather of the data annotation and its subjectivity. According to this, we can also confirm the difficulties of the SKIC data collections.

For the ACL dataset, we observe that the category *Language Resource (LR)* is mis-classified as *Other* class. For example, *treebank* is annotated as *LR* but our model predicts it as *Other*. Further, the data imbalanced also causes performance degradation in the ACL dataset. This is because the proportion of *LR*, which is one of the classes in the ACL dataset, is very small in size (5.7%), while the percentage of *Other* is significantly higher (42.8%). Interestingly, in contrast to the above two target datasets, in the SciIE dataset, our SciBERT transfer learning generally performs very well.

6 Conclusion, Discussion, and Future Work

We investigated the performance of data-rich intermediate task transfer learning for scientific keyphrase identification and classification (SKIC) using pre-trained language models BERT and SciBERT. We perform experiments on SciBERT and BERT with a total of 42 pairs of intermediate and target tasks, where intermediate tasks are drawn from sequence tagging and natural language inference. We found that employing sequence tagging tasks as intermediate tasks on SciBERT performs the best on three publicly available keyphrase identification and classification datasets. Further, the intermediate task transfer learning using SciBERT outperforms both the previous work and strong baselines by a large margin. Specifically, for the SemEval 2017 Task 10 dataset, using FrameNet as an intermediate task on SciBERT transfer learning yields +7.19 improvement in F1 in keyphrase identification, and using POS Tagging yields +8.69 improvement in F1 in keyphrase classification. For the ACL RD-TEC 2.0 dataset, using POS Tagging leads to +6.16 F1 in keyphrase identification, and +4.79 F1 in keyphrase classification. For the SciIE dataset, using Supersense brings +1.63 F1 in keyphrase identification, and +7.79 F1 in keyphrase classification. According to these results, syntactic structure learning related intermediate tasks such as POS Tagging are preferable for SKIC tasks. Further, looking at our intermediate task domain difference that lies in natural language inference tasks, we explore their impact on the pre-trained language models. In particular, we empirically showed that using MNLI as an intermediate task on SciBERT transfer learning returns higher performance than employing SciTail as an intermediate task on BERT transfer learning. Future works in this area will benefit from the improvement of the available intermediate tasks and other related intermediate tasks will be explored. Moreover, a better understanding of when and why these intermediate tasks are working is one of the interesting future directions.

Interestingly, we observe that BERT suffers from a serious drop in performance possibly due to catastrophic forgetting. In particular, we observe that almost all of the intermediate tasks fail to provide better starting points for SKIC pre-trained language models fine-tuning. While SemEval 2017 Task 10 dataset achieves the best results when transfer from Chunking on BERT, this result is lower than single SciBERT fine-tuning. On ACL RD-TEC 2.0 and SciIE, no intermediate task produces higher performance on BERT intermediate task transfer learning. One potential reason could be a large vocabulary gap between our domain and the collections used to pre-train BERT. In the future, we plan to analyze the differences between BERT and SciBERT to better understand the effects of transfer learning for SKIC.

Acknowledgements

We thank Isabelle Augenstein for several clarifications of the task and the evaluation approach. We also thank our anonymous reviewers for their constructive comments and feedback, which helped improve our paper. This research is supported in part by NSF CAREER award #1802358, NSF CRI award #1823292, and UIC Discovery Partners Institute to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

References

Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. 2019. Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2551–2557. ACM.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 341–346, Vancouver, Canada, July. Association for Computational Linguistics.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August. Association for Computational Linguistics.

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, Hong Kong, China, November. Association for Computational Linguistics.

Hung-Hsuan Chen, II Ororbia, G Alexander, and C Lee Giles. 2015. Expertseer: A keyphrase based expert recommender for digital libraries. *arXiv preprint arXiv:1511.02058*.

Wang Chen, Hou Pong Chan, Piji Li, and Irwin King. 2020. Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1095–1105, Online, July. Association for Computational Linguistics.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. 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 2089–2095, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56, March.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China, November. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China, November. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of twitter. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden, July. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy, July. Association for Computational Linguistics.

Lung-Hao Lee, Kuei-Ching Lee, and Yuen-Hsien Tseng. 2017. The NTNU system at SemEval-2017 task 10: Extracting keyphrases and relations from scientific publications using multiple conditional random fields. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 951–955, Vancouver, Canada, August. Association for Computational Linguistics.

Sijia Liu, Feichen Shen, Vipin Chaudhary, and Hongfang Liu. 2017. MayoNLP at SemEval 2017 task 10: Word embedding distance pattern for keyphrase classification in scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 956–960, Vancouver, Canada, August. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. 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*, pages 3219–3232, Brussels, Belgium, October-November. Association for Computational Linguistics.

Erwin Marsi, Utpal Kumar Sikdar, Cristina Marco, Biswanath Barik, and Rune Sætre. 2017. NTNU-1@ScienceIE at SemEval-2017 task 10: Identifying and labelling keyphrases with conditional random fields. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 938–941, Vancouver, Canada, August. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada, July. Association for Computational Linguistics.

Krutarth Patel and Cornelia Caragea. 2019. Exploring word embeddings in crf-based keyphrase extraction from research papers. In Mayank Kejriwal, Pedro A. Szekely, and Raphaël Troncy, editors, *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 37–44. ACM.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July. Association for Computational Linguistics.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Silvia Quarteroni and Suresh Manandhar. 2006. User modeling for adaptive question answering and information retrieval. In *FLAIRS Conference*, pages 776–781.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado, May–June. Association for Computational Linguistics.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia, July. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. 2019. Open domain web keyphrase extraction beyond language modeling. 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 5175–5184, Hong Kong, China, November. Association for Computational Linguistics.