Unsupervised Domain Adaptation for Keyphrase Generation using Citation Contexts

Anonymous ACL submission

Abstract

Adapting keyphrase generation models to new domains typically involves few-shot fine-tuning with in-domain labeled data. However, annotating documents with keyphrases is often prohibitively expensive and impractical, requiring expert annotators. This paper presents `silk`, an unsupervised method designed to address this issue by extracting silver-standard keyphrases from citation contexts to create synthetic labeled data for domain adaptation. Extensive experiments across three distinct domains demonstrate that our method yields high-quality synthetic samples, resulting in significant and consistent improvements in in-domain performance over strong baselines.

1 Introduction

Keyphrase generation aims at automatically predicting a set of keyphrases —words or phrases that represent the main concepts— given a source text. Because they distill the important information from documents, keyphrases are useful for many applications in natural language processing and information retrieval, most notably for document indexing (Fagan, 1987; Zhai, 1997; Jones and Staveley, 1999; Gutwin et al., 1999; Boudin et al., 2020) and summarization (Zha, 2002; Wan et al., 2007; Liu et al., 2021; Koto et al., 2022). Keyphrase generation differs from its extractive counterpart in that it requires the capability of predicting keyphrases that do not necessarily appear in the source text (Liu et al., 2011; Meng et al., 2017). Current models for this task are built upon sequence-to-sequence models, and achieve remarkable prediction performance when a large amount of labeled data is available (Meng et al., 2021).

However, keyphrase-labeled data is notably scarce even for resource-rich languages. To date, there are only a handful of available datasets large enough to train keyphrase generation mode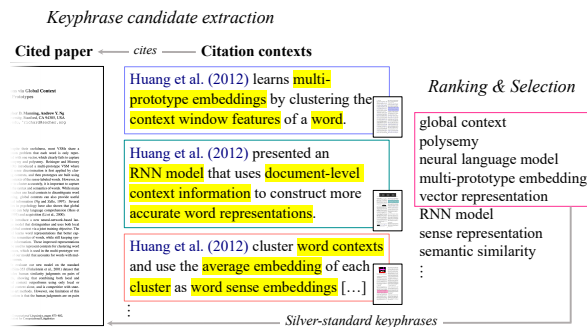ls, therefore restricting their applicability to specific domains (Ye and Wang, 2018; Wu et al., 2022; Garg et al., 2023). Here, we are concerned with generating keyphrases from scientific papers, for which datasets only exist in the broader scope of computer science (Meng et al., 2017; Mahata et al., 2022; Garg et al., 2022) and biomedicine (Houbre et al., 2022). This data scarcity issue is all the more important since current models demonstrate very limited generalization capabilities (Gallina et al., 2020; Meng et al., 2021). All of this, coupled with the high computational cost of training models, underscores the necessity of developing domain adaptation methods for keyphrase generation.

An effective strategy for addressing this challenge involves low-resource fine-tuning (Wu et al., 2022; Meng et al., 2023), wherein a pre-trained model is exposed to a limited amount of in-domain data with annotated keyphrases. Nevertheless, annotating even a limited number of documents can be prohibitively expensive, and often impractical due to the necessity for expert annotators (Chau et al., 2020). Finding a way to collect such data in an unsupervised fashion would open up possibilities for effortlessly adapting models to new domains. Here, we propose `silk`, a method to do so that relies on extracting silver-standard keyphrases from citation contexts to generate synthetic labeled data for domain adaptation (see Figure 1).

Figure 1: Illustration of the `silk` method for mining silver-standard keyphrases from citation contexts.

Citation contexts —text passages within the citing document containing the reference— often highlight the contributions of a cited paper, and have be shown to be useful not only for paper summarization (Nakov et al., 2004; Schwartz and Hearst, 2006; Mei and Zhai, 2008; Abu-Jbara and Radev, 2011; Mao et al., 2022, *inter alia*), but also for tasks such as claim verification (Wadden et al., 2020) or information extraction (Viswanathan et al., 2021). In this paper, we advocate for using citation contexts, specifically in the *mining of phrases representing the key concepts of cited papers*, to generate synthetic data for adapting keyphrase generation models to new domains. Earlier research on keyphrase extraction has emphasized the value of citation context information as a feature for ranking phrases (Das Gollapalli and Caragea, 2014; Caragea et al., 2014). We take this idea further and explore how it can be applied to create silver-labeled in-domain data for fine-tuning keyphrase generation models. Our contributions can be summarized as follows:

(1) We propose silk, a method that leverages citation contexts to create synthetic samples of documents paired with silver-standard keyphrases for adapting keyphrase generation models to new domains.

(2) We apply our method on three distinct scientific domains —namely, Natural Language Processing, Astrophysics and Paleontology—, thereby creating new adaptation data for each domain. We further provide three human-labeled test sets to assess the performance of keyphrase generation across these domains. We view this effort as a significant contribution of our work.

(3) We conduct experiments on few-shot fine-tuning a pre-trained model for keyphrase generation and report significant improvements in in-domain performance using synthetic samples generated by silk. Additionally, we undertake further experiments to validate the quality of the synthetic samples through both empirical (§5.1) and human (§5.3) evaluation, and we examine whether our adapted models experience catastrophic forgetting of the initial domain (§5.2).

Our code, model weights and data are available at http://anonymous.

2 Method

This section describes the implementation details of our method for producing synthetic fine-tuning data from citation contexts. Given a collection of in-domain scientific documents $\mathcal{D}$, we start by extracting the subset of sentences that contain citation anchors to build a set of citation contexts $\mathcal{C}$. Heuristics are applied to filter out citation contexts that either reference a document $d \notin \mathcal{D}$ or whose purpose of citing is ambiguous (i.e. containing multiple scattered citation anchors throughout the text). For each cited document $d \in \mathcal{D}$, we extract all the phrases[1] from its title t_d, abstract a_d and corresponding citation contexts c_d to build a set of silver-standard keyphrase candidates $\mathcal{P}_d$. Our method for generating synthetic samples from pairs of $(d, \mathcal{P}_d)$ involves three steps, which are described below.

Step 1: Ranking Keyphrase Candidates

We rank each keyphrase candidate $p \in \mathcal{P}_d$ based on three criteria:

- its salience, defined as the presence of p in t_d, a_d and c_d. Here, we assume that a phrase simultaneously occurring in all elements holds greater importance that a phrase found solely in one or two of them. A boosting parameter $\alpha = \{1, 1.5, 2\}$ is introduced to prioritize phrases based on the number of elements in which they appear.

- its relevance, computed as the cosine distance between the embedding vectors of p and t_d. We use the title as a high-level summary of the document, and assume that relevant phrases should be semantically close to it. We leverage SPECTER[2] (Cohan et al., 2020), a BERT-based model pre-trained on scientific documents, to compute the embedding vectors.

- its reliability, estimated by the number of citation contexts in which p occurs. We rely on the citation context frequency as a means to estimate how reliable a phrase is, the hypothesis being that phrases that appears in multiple citation contexts are more likely to be reliable. Specifically, we use the log-frequency of p in c_d to squash the range of values in a log-scale.

---

[1] We use spacy (en_core_web_sm model) and consider the noun phrases (/Adj*Noun+/) in their lemma forms as candidates. Irrelevant candidates are filtered out using a stoplist of high-frequency phrases.

[2] https://huggingface.co/sentence-transformers/allenai-specter

More formally, given a document d, the score of a keyphrase candidate p is calculated as follows:

$$\text{score}(p, d) = \alpha_{(p)} \cdot \text{cos-sim}\big(\text{emb}(p), \text{emb}(t_d)\big)$$
$$\cdot \log\big(\text{freq}_{\text{cc}}(p)\big) \quad (1)$$

where $\text{emb}(s)$ denotes the embedding vector output of SPECTER for input text s, $\text{freq}_{\text{cc}}(p)$ is the number of citation contexts in which p occurs.

Step 2: Selecting Silver-Standard Keyphrases

To select the optimal subset of phrases from $\mathcal{P}_d$, we define a set of constraints that mirrors the typical characteristics found in gold standard keyphrases of scientific papers. Building on past observations, our objective is to select between 3 and 5 phrases per document, comprising up to 3 phrases from its content (i.e. occurring in t_d or a_d) and the remainder from the citation contexts. We promote the selection of diverse keyphrases by introducing a maximum cross-phrase similarity threshold parameter λ_x. This parameter prevents the inclusion of redundant candidates, as determined by the cosine distance between their embedding vectors. Because candidates extracted from citation contexts are inherently noisy, we introduce a second threshold parameter λ_r to filter out spurious candidates based on their relevance scores. In our study, both parameters are determined based on their observed values in the validation split of the KP20k dataset (Meng et al., 2017); specifically, $\lambda_x = 0.85$ and $\lambda_r = 0.75$.

Step 3: Ordering Samples by Confidence

The final step involves ordering the cited documents based on how confident our method is in its silver-standard keyphrases, and selecting the top-N ranked documents as synthetic labeled data. Here, we determine the confidence of our method by averaging the scores of its silver-standard keyphrases, as computed in Equation 1. We remind that our objective is to generate small, high quality in-domain data for fine-tuning keyphrase generation models, which advocates for a conservative approach.

3   Datasets

We use the widely adopted KP20k dataset (Meng et al., 2017) as a starting point for pre-training keyphrase generation models. This dataset contains $\approx$ 514K scientific documents (titles and abstracts) paired with author-assigned keyphrases in the broader domain of computer science. We investigate the effectiveness of our domain adaptation method across three distinct scientific domains: Natural Language Processing (nlp), Astrophysics (astro), and Paleontology (paleo). These domains differ with increasing distances from the initial KP20k dataset, with nlp being the closest and paleo standing as the furthest. This section gives details about the data we use for each domain, presents the statistics of the resulting synthetic in-domain data we generate, and describes how we collect[3] annotated test data to validate the usefulness of our method for domain adaptation.

3.1   Natural Language Processing

For the nlp domain, we use the ACL Anthology Sentence Corpus[4] that contains the sentences of 65 662 papers from the ACL Anthology up until 2022. For quality reasons, we only consider sentences from papers published in the last 20 years (2003 and upwards) and occurring within the introduction and related work sections. From these, we extracted 260 324 citation contexts with the restriction that they include at least one citation to a paper within the ACL Anthology. For each cited paper, we applied our method to extract silver-standard keyphrases from citation contexts, resulting in a confidence-ordered list of 6 199 synthetic samples.

As most papers in the ACL Anthology do not provide keyphrases, we mainly relied on related conferences and journals to compile the test data for the nlp domain. More precisely, we manually collected a set of 212 documents (title and abstract) with author-assigned keyphrases from a variety of sources (e.g. LREC, SIGIR, CIKM).

3.2   Astrophysics

Astrophysics is a branch of astronomy concerned with the physical nature of stars, galaxies, interstellar clouds, and other objects. For the astro domain, we use the unarXive 2022 dataset (Saier et al., 2023) that contains 1.9M full-text papers from arXiv. We selected the subset of 198 349 papers that belong to the Astrophysics category (astro-ph), and extracted 133 320 citation contexts originating from the introduction sections of these papers. Applying our method for each cited paper results in a confidence-ordered list of 2 680 documents with silver-standard keyphrases.

---

[3]Detailed information on the sources can be found in A.4.
[4]https://kmcs.nii.ac.jp/resource/AASC/

For the `astro` test data, we manually collected a set of 255 documents (title and abstract) paired with author-assigned[5] keyphrases from both arXiv and journals. To guarantee diversity across various topics, we uniformly selected 20 documents from each astrophysics sub-category in arXiv and retrieved documents from broader-scope journals.

3.3 Paleontology

Palaeontology is the study of fossils and the evolution of life on Earth. To the best of our knowledge, no dataset of scientific papers is available for the `paleo` domain. To fill that gap, we collected 12 353 open- or free-access papers in PDF format from a wide range of journals in Paleontology.[6] We use GROBID[7] for extracting the full-text from PDF papers, detecting inline citations and parsing bibliography, as it was shown to outperform other freely available tools (Meuschke et al., 2023; Rohatgi et al., 2023). From the XML output of GROBID, we extracted 53 133 citation contexts from the introductory parts of the papers (i.e. "*Introduction*", "*Materials and Methods*" and "*Geological Settings*"). Applying our method on such a small collection yields only 130 documents with silver-standard keyphrases. To generate sufficient data for fine-tuning keyphrase generation models, we adjusted the threshold for candidate relevance (i.e. $\lambda_r = 0.75 \rightarrow 0.60$) and queried the Semantic Scholar API[8] to include cited papers not present in our collection. These modifications resulted in our method generating a confidence-ordered list of 2 806 documents with silver-standard keyphrases.

For the `paleo` test data, we manually collected a set of 244 documents, each paired with author-assigned keyphrases, sourced from approximately 10 different journals that encompass a wide spectrum of palaeontological topics (e.g. palaeogeography, palaeoecology or stratigraphy).

3.4 Statistics and Analysis

In this section, our aim is to deepen our understanding of the characteristics of the datasets we use for each domain and to assess how the compiled test data aligns with existing test datasets.

Table 1 summarizes the statistics of the datasets for each domain we apply our method on. There

is a noticeable diversity in characteristics across the datasets, with `nlp` showing the highest citation rate per document and `paleo` the lowest. We suspect there are two reasons for this. First, papers within the `nlp` domain seem to garner higher average citations compared to papers in the other two domains. Second, papers from `paleo` tend to cite works from both related domains (e.g. Biology, Geology) and sources outside our collection of gathered papers. Conversely, the average number of candidate keyphrases per document —those found in the title, abstract, or citation contexts— remains stable across the domains ($\approx$80 candidates).

	nlp	astro	paleo
# documents	65 662	198 349	12 353
# citation contexts	260 324	133 320	53 133
# cited doc	32 448	20 436	3 252
cites / doc	6.0	3.2	1.6
phrases / cited doc	72.9	76.8	87.4
↓ **silk** *(top-1K - all)*			*datasets* ↑
doc len. (tokens)	149	202	278
keyphrase / doc	3.9 3.6	3.6 3.5	3.6 3.6
keyphrase len.	1.8	1.9	1.6
% abs keyphrases	23.7 21.8	16.7 14.8	4.3 5.3

Table 1: Statistics for the datasets and the top-1K synthetics samples generated by `silk` for each domain.

Upon examining the synthetic fine-tuning data generated by our method (restricted to the top-1K), we observe that `nlp` documents are nearly half the length of those in the `paleo` domain, while `astro` documents fall in-between. These differences in length directly impact the ratio of absent keyphrases[9], decreasing from 24% to below 10%. These numbers further decrease when computed beyond the top-1K, as the number of citation contexts declines and, consequently, as the pool of absent keyphrase candidates reduces. Constraints we introduced for selecting the optimal subset of phrases allow for an average of about 4 silver keyphrases per document, predominantly unigrams and bigrams, which is in line with both past observations and the test data we compiled (see Table 2).

To analyze the disparities between the domains we selected, and also how they depart from KP20k (initial domain) and from other existing test datasets for keyphrase generation, we compare the

Dataset	#doc	len$_{doc}$	#kp	len$_{kp}$	%abs
KP20k	20K	176	5.2	2.0	41.5
nlp	212	210	4.1	2.0	36.7
astro	255	224	4.9	2.1	47.8
paleo	244	255	5.5	1.5	38.6
Inspec	500	134	9.8	2.3	21.4
NUS	211	182	11.7	2.1	45.2
SemEval	100	203	14.5	2.1	60.7

Table 2: Statistics for the test data we collected for each domain in comparison with commonly used test sets.

Figure 2: t-SNE 2-D projections of the gold keyphrases from ● KP20k, ● nlp, ● astro and ● paleo. We leverage SPECTER to compute the keyphrase embeddings and use the first 500 documents from KP20k for clarity.

main statistics of their test splits in Table 2. Here, we include three additional datasets, Inspec (Hulth, 2003), NUS (Nguyen and Kan, 2007) and SemEval-2010 (Kim et al., 2010), that are composed of scientific abstracts in the computer science domain. Together with KP20k, these are probably the most commonly-used datasets for evaluating keyphrase generation models. Overall, we observe many similarities between KP20k and the test data we collected for each domain, whether in terms of the number of gold keyphrases ($\approx$5 per document), their average length ($\approx$2 tokens) or the ratio of absent keyphrases ($\approx$40%). This suggests a uniform trend in author-assigned keyphrases across scientific domains, which should facilitate generalization for keyphrase generation models. It should be noted that higher number of gold keyphrases in NUS, SemEval-2010 and Inspec stems from their distinct annotation processes, with the former two combining author- and reader-assigned keyphrases and the latter relying on professional indexers. Comparing the sizes of our domain-specific test data with those of the test splits in existing datasets shows that they are on a similar scale.

Lastly, we examine the differences between the domains from a semantic perspective. Figure 2 shows a t-SNE visualization (van der Maaten and Hinton, 2008) of the gold keyphrases in the test data that we collected for each domain and those of the KP20k test split. We clearly discern the different domains within the vector space, roughly dividing it into four clusters. The most notable overlap occurs between nlp and KP20k (computer science), whereas astro and paleo exhibit clear separation. These visual insights support our initial assumptions regarding the growing differences of our selected domains from KP20k, with nlp being the closest and paleo standing as the furthest.

4 Experimental Settings

4.1 Initial Model

We use BART (Lewis et al., 2020) as our initial pre-trained language model and perform fine-tuning on the KP20k training set for 15 epochs, following (Wu et al., 2023). BART was shown to yield state-of-the-art performance in keyphrase generation (Zhao et al., 2022; Meng et al., 2023), surpassing other pre-trained language models, such as T5 (Wu et al., 2023). Following previous work, we fine-tune BART in a ONE2MANY setting (Yuan et al., 2020), that is, given a source text as input, the task is to generate keyphrases as a single sequence of delimiter-separated phrases. During fine-tuning, gold keyphrases are arranged in the present-absent order which was found to give the best results (Meng et al., 2021). At test time, we use either greedy decoding and let the model generate the most probable keyphrases, or beam search (K=20) and assemble the top-k keyphrases from all the beams as the model output. Implementation details are given in Appendix A.2.

4.2 Domain Adaptation

For adapting our fine-tuned BART model to a specific domain, we continue fine-tuning it on the synthetic labeled data generated by silk for 3 epochs. Specifically, we use the top-N most confident silver-labeled examples to further fine-tune BART, creating three gradually adapted models for each domain by varying $N \in \{500, 1K, 2K\}$. We compare the effectiveness of our domain adaptation method with that of the only other unsupervised approach we are aware of, which is self-learning (Ye and Wang, 2018; Meng et al., 2023). Self-learning

406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452

consists in using a model to generate pseudo-labels for in-domain documents and then re-train itself on this data. Here, we use our fine-tuned BART model to generate keyphrases for the same documents as those produced by `silk`, and further fine-tune it on this self-labeled data for 3 epochs.

4.3 Baselines

Although the focus of this work is domain adaptation for keyphrase generation, we also provide the results of two baselines as a point of reference. The first baseline is MultiPartiteRank (Boudin, 2018), an unsupervised method for keyphrase extraction that leverages graph-based ranking and topical information. Despite being limited to present keyphrases, MultiPartiteRank yields the best results among non deep learning methods (Do et al., 2023). We use the author's implementation provided by the pke toolkit.[10] The second baseline is One2Set (Ye et al., 2021), a Transformer-based model for keyphrase generation that utilizes learned control codes to generate a set of keyphrases. This model achieves strong performance, often on-par with state-of-the-art models (Zhao et al., 2022; Wu et al., 2023). We use the model weights released by the authors, which were trained on KP20k.[11]

4.4 Datasets and Evaluation Metrics

We use the test split of KP20k for evaluating the initial performance of the models, and our manually collected test sets to assess their in-domain performance. Detailed statistics for these datasets are presented in Table 2. Following common practice, we evaluate the performance of the models in terms of F_1 score using exact match between gold and predicted keyphrases. Stemming (Porter stemmer) is applied to reduce the number of mismatches and duplicates are removed. We compute the scores both at the top-k predicted keyphrases with $k \in 5, 10$, and at the number M of keyphrases predicted by the models as proposed in (Yuan et al., 2020). For $F_1@k$ scores, if the number of predicted keyphrases is below k, we append incorrect predictions until it reaches exactly k keyphrases. We also report scores for present and absent keyphrases separately to get more insights about the extractive and generative capabilities of the models. We compute the Student's paired t-test to assess the statistical significance of our results at $p < 0.05$.

---

[10] https://github.com/boudinfl/pke
[11] https://github.com/jiacheng-ye/kg_one2set

4.5 Performance of Models on KP20k

453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490

Table 3 presents the results of our fine-tuned BART model (hereafter denoted as BART-FT) and the two baselines on the test split of KP20k. It should be noted that MultiPartiteRank cannot be assessed using $F_1@M$ as it requires setting a top-k parameter, and that One2Set cannot be assessed using $F_1@10$ since it only outputs the most probable keyphrases ($\approx$7 per document). Overall, BART-FT demonstrates superior performance, significantly outperforming the baselines for both all and only the present keyphrases. We observe that One2Set achieves the best scores for the absent keyphrases, confirming previous findings (Wu et al., 2023). Despite employing the model weights and code provided by the authors, our scores for One2Set are lower than those reported in the original paper. This is due to preprocessing discrepancies. For instance, in (Ye et al., 2021), abbreviations are removed from gold keyphrases and digits are replaced with the special token while we retain them in their original form. In light of these results, we argue that BART-FT is a strong model for keyphrase generation, providing a solid basis for the application of our domain adaptation method.

Model	$F_1@M$			$F_1@5$			$F_1@10$		
	all	pres	abs	all	pres	abs	all	pres	abs
MPRank	-			13.6	17.3	-	12.8	15.1	-
One2Set	23.2	35.1	5.5†	23.5	29.9	4.2	-		
BART-FT	28.7†	37.3†	2.4	28.0†	35.5†	5.9†	25.4†	29.2†	5.8

Table 3: Performance comparison on the KP20k test set, with † indicating statistical significance. Scores for present and absent keyphrases separately are reported.

5 Results

Table 4 presents the results of the keyphrase generation models and our domain adaptation method on each domain.[12] We observe that `silk` brings consistent and significant improvements over BART-FT on the three domains. The best overall performance is achieved by fine-tuning the model with the top-1K most confident synthetic samples, however gains are observed with just the top-500 samples. Self-learning for domain adaptation yields only marginal gains at best and often degrades performance. This suggests that the initial performance of BART-FT on these domains is not

---

[12] See Table 13 in Appendix A for the detailed results of present and absent keyphrases.

Model	FT	nlp			astro			paleo		
		$F_1@M$	$F_1@5$	$F_1@10$	$F_1@M$	$F_1@5$	$F_1@10$	$F_1@M$	$F_1@5$	$F_1@10$
MPRank		-	13.3	11.9	-	12.8	10.8	-	13.1	13.1
One2Set		21.6	24.1	-	13.2	12.7	-	13.4	12.1	-
BART-FT		30.8	29.8	24.8	19.1	18.6	16.5	18.4	18.9	18.8
+self-learning	500	31.2	30.0	24.5	19.1	18.9	16.2	18.8	18.7	18.6
	1K	30.6	30.7	24.1	19.6	19.5	16.5	19.3	19.2	18.8
	2K	29.8	29.2	24.4	18.4	19.1	16.3	18.9	19.7	18.4
+silk (ours)	500	31.0	29.7	**25.2**	18.8	19.2	17.0	19.0	19.3	19.2
	1K	**33.7**†	**32.2**†	**25.2**	19.2	20.4†	17.7†	**19.4**	**20.4**†	**19.5**
	2K	29.8	29.2	24.4	**20.3**	21.5†	**17.9**†	17.7	19.1	18.0

Table 4: Performance of keyphrase generation models on the nlp, astro and paleo domains for all keyphrases (i.e. present and absent combined). Values in **bold** indicate best scores and † indicates significance over BART-FT.

sufficient to generate high-quality pseudo-labels. A closer look at the numbers shows that BART-FT performs comparably on nlp as it does on KP20k (see Table 3), but it gives substantially lower scores on paleo and astro. This empirically confirms the growing distance between KP20k and these three domains, correlating model performance with the distance from the initial domain.

Among the three domains, paleo poses the greatest challenge for our method. We see two main reasons for this. First, the limited size of our collection of full-text papers ($\approx$12K), and the necessary parameter adjustments made to accommodate it, adversely affect the quality of the synthetic samples. Second, the paleo domain in itself appears to be more challenging to handle due to its interdisciplinary nature, spanning subjects such as geology, biology, history, and ecology, among others. Examining the performance of the baselines, we observe the poor generalization of One2Set whose results nearly drop by half for non-computer science domains, and that is even surpassed by MultiPartiteRank. This latter delivers consistent, albeit modest, performance across domains which makes it relevant as an estimator of lower-bound performance for research on domain adaptation.

In the remainder of this section, we conduct additional experiments to 1) validate the confidence ranking of the synthetic samples in silk; 2) examine whether our domain adaptation method causes catastrophic forgetting of the initial domain; and 3) showcase the quality of the silver-standard keyphrases generated by silk through human evaluation.

5.1 Confidence Ranking of synthetic samples

The purpose of silk is to generate small, high quality in-domain data for fine-tuning keyphrase generation models. Accordingly, synthetic samples are ordered by confidence (described in §2) and only the top-N ranked samples are employed for adapting models. To validate the quality of our ranking, and consequently the effectiveness of our keyphrase candidate scoring function (see Equation 1), we compare the performance of BART-FT when we continue the fine-tuning with the top-1K, bottom-1K and a random selection of 1K samples. Results are presented in Table 5. We note that, uniformly across the three domains, the random and top-1K ordering schemes lead to improvements, with top-1K yielding the best results. In contrast, using the least confident samples (bottom-1K) systematically degrades the performance. Insights from these results are twofold: 1) our confidence ranking proves to be beneficial for selecting high-quality synthetic samples, and 2) even samples beyond the top-1K are qualitative enough for domain adaptation.

Model	nlp		astro		paleo	
	$F_1@M$	$F_1@10$	$F_1@M$	$F_1@10$	$F_1@M$	$F_1@10$
BART-FT	30.8	24.9	19.1	16.5	18.4	18.8
+silk (top)	33.7†	25.2	19.2	17.7†	19.4	19.5
+silk (ran)	33.2†	25.1	19.3	17.4	19.4	19.4
+silk (bot)	27.8	21.4	16.5	15.1	16.9	17.3

Table 5: Performance of BART-FT fine-tuned on the top-1K, bottom-1K and random-1K (averaged over 5 runs) samples. † indicates significance over BART-FT.

5.2 Forgetting of Domain Adaptation

Although continued training is effective for domain adaptation, it has been found to adversely affect performance in the initial domain for language generation tasks (Thompson et al., 2019; Li et al., 2022). Here, we investigate whether this phenomenon, referred to in the literature as catastrophic forgetting (French, 1999), also manifests in our adapted models. Table 6 presents the results of our domain adapted BART-FT models (using 1K synthetic samples) on the KP20k test set. Overall, we observe no drop in performance for our adapted models. Rather surprisingly, we notice small improvements in $F_1@k$ scores over the initial BART-FT model. Upon closer examination, these gains derive from improved extractive capabilities, while the scores for absent keyphrases consistently degrade. We hypothesise that the domain adaption process makes the model lose generative ability and reinforces its extractive capability which translates more effectively across domains.

Model	$F_1@M$			$F_1@5$			$F_1@10$		
	all	pres	abs	all	pres	abs	all	pres	abs
BART-FT	28.7	37.3	2.4	28.0	35.5	5.9	25.4	29.2	5.8
+silk (nlp)	28.6	37.5	1.6	28.3	35.9	5.5	25.7	29.7	5.4
+silk (astro)	28.7	37.8	1.7	28.7	36.4	5.9	25.9	29.8	5.9
+silk (paleo)	28.4	37.5	1.4	28.6	36.2	5.7	25.8	29.7	5.6

Table 6: Performance comparison of BART-FT and its adaptions (silk 1K) on the KP20k test set. Scores for present and absent keyphrases are also reported.

5.3 Qualitative analysis

We further examine the quality of the synthetic samples produced with silk by conducting a manual evaluation of the top-100 samples of the nlp domain.[13] Annotators were instructed to assess the relevance of silver-standard keyphrases using a 3-point scale: "*not relevant*", "*partially relevant*" and "*relevant*". Additionally, we requested annotators to assess the well-formedness of the keyphrases with a binary rating. To quantify the qualitative difference between silk keyphrases and automatically generated ones, we perform a second round of human evaluation for BART-FT utilizing the same top-100 samples. Table 7 presents the results of our qualitative analysis. First, we note that nearly all silk keyphrases are well-formed, with any exceptions attributable to tagging errors (e.g. "*inter*

---
[13]Annotation guidelines and examples can be found in A.3.

alia"). More importantly, we observe that 80% of silk keyphrases are relevant, demonstrating the effectiveness of our method. In contrast, only 54.5% of the keyphrases generated by BART-FT are deemed relevant, which explains why the self-learning approach to domain adaptation falls short. We also note that BART-FT tends to generate more keyphrases ($\approx$5.5 per document), many of which are broader terms that are often irrelevant for the NLP domain (e.g. "*natural language processing*", "*statistics*" or "*machine learning*").

Model	#kp	WFness		Relevance		
		no	yes	no	part.	yes
BART-FT	545	6.2	93.8	35.0	10.5	54.5
silk	411	2.9	97.1	11.9	8.0	80.0

Table 7: Human evaluation results (%) in terms of well-formedness and relevance of the top-100 nlp samples generated by silk and re-annotated using BART-FT.

6 Conclusion and Future Work

In this paper, we propose silk, an unsupervised method that relies on citation contexts to create silver-standard data for adapting keyphrase generation models to new domains. We conduct experiments across three distinct scientific domains and demonstrate the effectiveness of our method for domain adaptation by few-shot fine-tuning a pre-trained model for keyphrase generation. We further validate the quality of the synthetic samples created by silk through human evaluation and analysis.

Our work addresses the issue of domain adaptation in keyphrase generation by introducing a solution that leverages citation contexts. Considering that citing papers is the *de-facto* means for discussing past work in scientific writing, we argue that it is possible to generate silver-standard data for most domains, provided that there is a minimal number of papers available. Such data would not only be useful for adapting existing models to new domains but also for keeping them up-to-date, given the rapid expansion of scientific literature and the evolving terminology across all domains. Mining silver keyphrases from citation contexts also adds another dimension in that they capture the contributions of a paper according to its peers. These do not necessarily align with the authors' keyphrases, raising further questions as to how pessimistic our intrinsic evaluation scores are.

Limitations

While our proposed method is both straightforward and effective, it is important to acknowledge its limitations. First, as our method leverages citation contexts, the generated synthetic samples are inherently biased towards highly cited papers and their corresponding research topics. Second, we did not optimize each component of our method, relying instead on heuristics for selecting and filtering citation contexts and scoring silver-standard keyphrases using a simple combination of criteria.

Since our work focuses on generating synthetic data for domain adaptation, and we did not search for the optimal fine-tuning parameters, and also relied on a single pre-trained model (BART-base). Even though we have conducted extensive experiments across three domains, it remains unclear how our findings generalize to other or larger pre-trained models. Manually evaluating the quality of automatically generated keyphrases is inherently subjective. Although we developed simple and detailed guidelines to minimize variability in assessments, it remains unclear how the results from our qualitative analysis extend beyond the top 100 samples in nlp or to the other two domains.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509, Portland, Oregon, USA. Association for Computational Linguistics.

Wasi Ahmad, Xiao Bai, Soomin Lee, and Kai-Wei Chang. 2021. Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. 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 1389–1404, Online. Association for Computational Linguistics.

Hareesh Bahuleyan and Layla El Asri. 2020. Diverse keyphrase generation with neural unlikelihood training. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5271–5287, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 667–672, New Orleans, Louisiana. Association for Computational Linguistics.

Florian Boudin and Ygor Gallina. 2021. Redefining absent keyphrases and their effect on retrieval effectiveness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4185–4193, Online. Association for Computational Linguistics.

Florian Boudin, Ygor Gallina, and Akiko Aizawa. 2020. Keyphrase generation for scientific document retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1126, Online. Association for Computational Linguistics.

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.

Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. 2020. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain. Association for Computational Linguistics.

Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. Keyphrase generation with correlation constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium. Association for Computational Linguistics.

Wang Chen, Yifan Gao, Jiani Zhang, Irwin King, and Michael R. Lyu. 2019. Title-guided encoding for keyphrase generation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

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, Online. Association for Computational Linguistics.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Lam Do, Pritom Saha Akash, and Kevin Chen-Chuan Chang. 2023. Unsupervised open-domain keyphrase generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10614–10627, Toronto, Canada. Association for Computational Linguistics.

J. Fagan. 1987. Automatic phrase indexing for document retrieval. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '87, page 91–101, New York, NY, USA. Association for Computing Machinery.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2020. Large-scale evaluation of keyphrase extraction models. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 271–278, New York, NY, USA. Association for Computing Machinery.

Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. 2022. Keyphrase generation beyond the boundaries of title and abstract. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5809–5821, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. 2023. Data augmentation for low-resource keyphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8442–8455, Toronto, Canada. Association for Computational Linguistics.

Carl Gutwin, Gordon Paynter, Ian Witten, Craig Nevill-Manning, and Eibe Frank. 1999. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, 27(1):81–104.

Maël Houbre, Florian Boudin, and Beatrice Daille. 2022. A large-scale dataset for biomedical keyphrase generation. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 47–53, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xiaoli Huang, Tongge Xu, Lvan Jiao, Yueran Zu, and Youmin Zhang. 2021. Adaptive beam search decoding for discrete keyphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13082–13089.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Steve Jones and Mark S. Staveley. 1999. Phrasier: A system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 160–167, New York, NY, USA. Association for Computing Machinery.

Jihyuk Kim, Myeongho Jeong, Seungtaek Choi, and Seung-won Hwang. 2021. Structure-augmented keyphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2657–2667, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. LipKey: A large-scale news dataset for absent keyphrases generation and abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3427–3437, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. 2022. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454, Seattle, United States. Association for Computational Linguistics.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2021. Keyword-aware abstractive summarization by extracting set-level intermediate summaries. In *Proceedings of the Web Conference 2021*, WWW '21, page 3042–3054, New York, NY, USA. Association for Computing Machinery.

Zhiyuan Liu, Xinxiong Chen, Yabin Zheng, and Maosong Sun. 2011. Automatic keyphrase extraction by bridging vocabulary gap. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 135–144, Portland, Oregon, USA. Association for Computational Linguistics.

10

Debanjan Mahata, Navneet Agarwal, Dibya Gautam, Amardeep Kumar, Swapnil Parekh, Yaman Kumar Singla, Anish Acharya, and Rajiv Ratn Shah. 2022. Ldkp: A dataset for identifying keyphrases from long scientific documents.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio. Association for Computational Linguistics.

Rui Meng, Tong Wang, Xingdi Yuan, Yingbo Zhou, and Daqing He. 2023. General-to-specific transfer labeling for domain adaptable keyphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1602–1618, Toronto, Canada. Association for Computational Linguistics.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–5007, Online. 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. Association for Computational Linguistics.

Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. 2023. A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pages 383–405, Cham. Springer Nature Switzerland.

Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, volume 4, pages 81–88. Citeseer.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Tarek Saier, Johan Krause, and Michael Färber. 2023. unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 66–70, Los Alamitos, CA, USA. IEEE Computer Society.

Ariel S. Schwartz and Marti Hearst. 2006. Summarizing key concepts using citation sentences. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 134–135, New York, New York. Association for Computational Linguistics.

Xianjie Shen, Yinghan Wang, Rui Meng, and Jingbo Shang. 2022. Unsupervised deep keyphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11303–11311.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the citation graph for scientific information extraction. 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 719–731, Online. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic. Association for Computational Linguistics.

11

Di Wu, Wasi Ahmad, and Kai-Wei Chang. 2023. Rethinking model selection and decoding for keyphrase generation with pre-trained sequence-to-sequence models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6642–6658, Singapore. Association for Computational Linguistics.

Di Wu, Wasi Ahmad, Sunipa Dev, and Kai-Wei Chang. 2022. Representation learning for resource-constrained keyphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 700–716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hai Ye and Lu Wang. 2018. Semi-supervised learning for neural keyphrase generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4142–4153, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Ye, Tao Gui, Yichao Luo, Yige Xu, and Qi Zhang. 2021. One2Set: Generating diverse keyphrases as a set. 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 4598–4608, Online. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2020. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975, Online. Association for Computational Linguistics.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 113–120, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai. 1997. Fast statistical parsing of noun phrases for document indexing. In *Fifth Conference on Applied Natural Language Processing*, pages 312–319, Washington, DC, USA. Association for Computational Linguistics.

Guangzhen Zhao, Guoshun Yin, Peng Yang, and Yu Yao. 2022. Keyphrase generation via soft and hard semantic corrections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7757–7768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jing Zhao and Yuxiang Zhang. 2019. Incorporating linguistic constraints into keyphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5224–5233, Florence, Italy. Association for Computational Linguistics.

A Appendices

A.1 Related work

Keyphrase generation was first introduced by (Liu et al., 2011) and subsequently formulated as a sequence-to-sequence language generation task by (Meng et al., 2017). They proposed an RNN-based encoder-decoder model with attention and copy mechanisms, which was later enhanced by the addition of decoding constraints to improve keyphrase diversity (Chen et al., 2018; Zhao and Zhang, 2019; Bahuleyan and El Asri, 2020; Yuan et al., 2020; Huang et al., 2021), or by learning to encode the structural information of input documents (Ye and Wang, 2018; Chen et al., 2019; Kim et al., 2021). Later work switched to Transformers-based models (Meng et al., 2021; Ye et al., 2021; Ahmad et al., 2021), reporting better performance. Recently, pre-trained language models (PLMs) have been used for keyphrase generation, predominantly through continued fine-tuning (Zhao et al., 2022; Meng et al., 2023; Wu et al., 2023).

Our work also intersects with unsupervised models for keyphrase generation (Shen et al., 2022; Do et al., 2023), which evaluate the informativeness of keyphrases based on their semantic similarity to the source document. Another direction to mitigate the data scarcity issue in keyphrase generation involves leveraging both labeled and unlabeled data for training. Ye and Wang (2018) proposed a self-learning approach to augment the training data with synthetic samples. Similarly, Meng et al. (2023) extended this concept to adapt models to new domains by generating domain-specific synthetic samples. In a low-resource setting, Garg et al. (2023) introduced a data augmentation method that leverages the full text of the documents to add diversity to the training samples.

Our work is closely related to the use of citation contexts in automated models for producing keyphrases. For keyphrase extraction, Das Gollapalli and Caragea (2014) proposed a graph-ranking approach that leverages citation contexts while scoring candidates, and Caragea et al. (2014) use the occurrence of candidates in citation contexts as a feature in a supervised model. For keyphrase generation, Garg et al. (2022) proposed to append citation contexts to enrich the input document.

A.2 Implementation Details

We use the BART-base model weights as our initial pre-trained language model and perform fine-tuning on the KP20k training set for 15 epochs. We use the AdamW optimizer with a learning rate of 1e-5 and a batch size of 24. Fine-tuning the model using 2 Nvidia GeForce RTX 2080 took 62 hours.

For adapting BART-FT to a each domain, we continue fine-tuning on $N \in \{500, 1K, 2K\}$ synthetic samples for 3 epochs. We use the AdamW optimizer with a learning rate of 1e-6 and a batch size of 16. Few-shot fine-tuning, conducted on a MacBook Pro M1 Max, required an average of 5 minutes per model, totaling 3 hours for all 12 models per domain.

A.3 Guidelines for manual evaluation

We evaluate the silver-standard keyphrases created by silk and those generated by BART-FT along two criteria: their relevance with respect to the source document, and their well-formedness. Annotators (authors of this paper) were given the title, the abstract and access to the full-text paper when evaluating the quality of the keyphrases. We perform manual evaluation on the top-100 synthetic samples generated by silk, confined to the nlp domain for which annotators have expertise.

Relevance is assessed on a 3-point scale, where 0 indicates that the keyphrase is not relevant, 1 that it is partially relevant (i.e. covering a related concept) and 2 that it is relevant to the source document.

Well-formedness is assessed on a binary scale, with 0 indicating that the keyphrase lacks proper form, such as being improperly structured (e.g. "*algorithms and data structures*") or not forming a self-contained phrase (e.g. "*large amount*"), while 1 signifies that the keyphrase is well-formed.

Orthographic variants occurring in a set of keyphrases (e.g. "*co-reference resolution*" and "*coreference resolution*") are identified, and only one of them is considered as relevant. We do not consider abbreviations as variants of their expanded forms. Broader terms such as "*natural language processing*", "*machine learning*" or "*neural networks*" are generally considered as too generic and not relevant.

An example of output for silk and BART-FT is shown is Table 8.

Get To The Point: Summarization with Pointer-Generator Networks (Bibkey: see-etal-2017-get)

Neural sequence-to-sequence models have provided a viable new approach for abstractive text summarization (meaning they are not restricted to simply selecting and rearranging passages from the original text). However, these models have two shortcomings: they are liable to reproduce factual details inaccurately, and they tend to repeat themselves. In this work we propose a novel architecture that augments the standard sequence-to-sequence attentional model in two orthogonal ways. First, we use a hybrid pointer-generator network that can copy words from the source text via pointing, which aids accurate reproduction of information, while retaining the ability to produce novel words through the generator. Second, we use coverage to keep track of what has been summarized, which discourages repetition. We apply our model to the CNN / Daily Mail summarization task, outperforming the current abstractive state-of-the-art by at least 2 ROUGE points.

silk	summarization, pointer-generator network, sequence-to-sequence model, copy mechanism, coverage mechanism
	well-formedness: 1 1 1 1 1 relevance: 1 1 1 1 1
BART-FT	summarization, sequence-to-sequence models, attentional models, cnn, daily mail, neural networks, text mining
	well-formedness: 1 1 1 1 1 1 1 relevance: 1 1 1 1 1 0 0

Improving Neural Machine Translation Models with Monolingual Data (Bibkey: sennrich-etal-2016-improving)

Neural Machine Translation (NMT) has obtained state-of-the art performance for several language pairs, while only using parallel data for training. Target-side monolingual data plays an important role in boosting fluency for phrase-based statistical machine translation, and we investigate the use of monolingual data for NMT. In contrast to previous work, which combines NMT models with separately trained language models, we note that encoder-decoder NMT architectures already have the capacity to learn the same information as a language model, and we explore strategies to train with monolingual data without changing the neural network architecture. By pairing monolingual training data with an automatic back-translation, we can treat it as additional parallel training data, and we obtain substantial improvements on the WMT 15 task English German (+2.8-3.7 BLEU), and for the low-resourced IWSLT 14 task Turkish->English (+2.1-3.4 BLEU), obtaining new state-of-the-art results. We also show that fine-tuning on in-domain monolingual and parallel data gives substantial improvements for the IWSLT 15 task English->German.

silk	neural machine translation, monolingual data, back-translation, data augmentation, synthetic parallel corpus
	well-formedness: 1 1 1 1 1 relevance: 1 1 1 1 1
BART-FT	neural machine translation, monolingual data, language models, back-translation, language modeling and translation, parallel training data
	well-formedness: 1 1 1 1 0 1 relevance: 1 1 1 1 0 0.5

Table 8: Examples of document (title and abstract) from the `nlp` domain with silver-standard keyphrases generated by `silk` and automatically generated keyphrases from BART-FT.

A.4 Sources used for collecting test data

Source	Session / Volume	#nb
SIGIR 2023	Language Models	6
	Question Answering	3
	Summarization & Text Generation	5
	Short Research Papers✋	16
CIKM 2023	Natural Language	24
WSDM 2023	Language Models and Text Mining	6
SIGIR 2022	NLP and Semantics	8
	Question Answering	4
	Sentiment Analysis and Classification	5
	Short Research Papers✋	14
CHI 2022	Natural Language	5
LREC 2022	Oral sessions✋	81
NLP Journal[14]	Volumes 2-5	35
	Total	212

Table 9: Detailed information on the sources of the test documents for the `nlp` domain. ✋ indicates that we manually selected the documents to filter out out-of-domain ones.

Source	Category / Year	#nb
arXiv (oct→dec 2023)	astro-ph.HE (High Energy Astro. Phenomena)	20
	astro-ph.CO (Cosmology and Nongalactic Astro.)	20
	astro-ph.IM (Instrumentation and Methods for Astro.)	20
	astro-ph.SR (Solar and Stellar Astro.)	20
	astro-ph.EP (Earth and Planetary Astro.)	20
	astro-ph.GA (Astro. of Galaxies)	20
Frontiers in Astro. Astrophysics	2022-23 (selected using arXiv keywords)	76
	2022-23	59
	Total	255

Table 10: Detailed information on the sources of the test documents for the `astro` domain.

Source	Year	#nb
Palaeontologia Electronica	2023-24	21
Acta Palaeontologica Polonica	2023	22
Palaeontology	2023	26
Cretaceous Research	2024	20
Palaeogeography, Palaeoclimatology, Palaeoecology	2024	47
Papers in Palaeontology	2023	29
Proc. Royal Soc. B: Biological Sciences	2023	25
Biology Letters	2023	25
Palaeobiodiversity and Palaeoenvironments	2023	16
	Total	244

Table 11: Detailed information on the sources of the test documents for the `paleo` domain.

Source	Licence	Year	#nb
Acta Geologica Sinica	open	2021-2023	21
Acta Palaeontologica Polonica	open	2002-2023	1 398
Alcheringa: An Australasian Journal of Palaeontology	open	2016-2023	32
Carnets de Geologie	open	2015-2023	147
Cretacious Research	open	2019-2023	81
Journal of paleontology	open	2015-2023	144
Journal of Systematic Palaeontology	open	2016-2023	34
Journal of Vertebrate Paleontology	open	2013-2023	95
Lethaia	open/free	2018-2023	239
Nature	open	2010-2023	705
Palaeobiodiversity and Palaeoenvironments	open	2002-2023	811
Palaeodiversity	open	2016-2023	74
Palaeogeography, Palaeoclimatology, Palaeoecology	open	2019-2023	201
Palaeontologia Electronica	open	1998–2023	841
Palaeontology	open/free	1999-2023	1 474
Paleobiology	open	2013-2023	133
Paleoceanography and Paleoclimatology	open	2014-2023	128
PalZ	open/free	2009-2023	651
Papers in Palaeontology	open	2015-2023	71
Plos Paleontology	open	2011-2017	237
Proceedings of the Royal Society B: Biological Sciences (paleontology)	open/free	2009-2023	354
PubMedfreef ulltext (query="Paleontology[MeSH Terms]")	open/free	1955-2023	3 462
Research in Paleontology and Stratigraphy	open	2019-2023	157
Royal Society Science (paleontology)	open/free	2014-2023	270
Royal Society Biology Letters (paleontology)	open/free	2009-2023	235
Swiss Journal of Palaeontology	open/free	2011-2023	282
Trends in Ecology and Evolution (paleobiology)	open	2020-2022	15
Zookeys (paleontology)	open	2015-2023	61
		Total	12 353

Table 12: Detailed information on the sources of the scientific papers collected for the Paleontology corpus.

Model	FT	nlp $F_1@M$		$F_1@5$		$F_1@10$		astro $F_1@M$		$F_1@5$		$F_1@10$		paleo $F_1@M$		$F_1@5$		$F_1@10$	
		pres	abs	pres	abs	pres	abs	pres	abs	pres	abs	pres	abs	pres	abs	pres	abs	pres	abs
MPRank		-		16.3	-	13.5	-	-		17.0	-	13.1	-	-		16.5	-	15.0	-
One2Set		36.1	**3.3**	28.3	2.8	-		20.1	**1.6**	17.3	1.2	-		18.6	0.3	16.5	0.2	-	
BART-FT		38.8	2.0	36.8	**3.7**	27.9	**3.6**	26.7	0.2	25.6	**1.4**	20.2	**1.4**	23.5	0.5	24.3	**1.3**	22.6	**1.2**
+self-learning	500	38.9	2.0	36.9	3.0	27.7	3.0	26.5	0.0	26.0	0.7	19.9	0.7	24.1	0.5	23.8	1.2	22.6	1.1
	1K	38.2	2.0	36.8	3.1	27.6	3.0	27.0	0.0	26.1	0.5	20.5	0.5	24.5	0.2	24.8	0.6	22.9	0.7
	2K	37.2	2.6	36.8	3.2	27.8	3.0	25.3	0.2	25.3	0.5	20.2	0.7	24.3	0.2	25.7	0.6	22.7	0.6
+silk (ours)	500	39.1	0.5	36.2	2.8	**28.2**	2.9	26.4	0.2	26.7	1.1	21.1†	1.1	24.4	0.5	24.8	1.0	**23.5**	0.9
	1K	**41.7**†	1.2	**38.3**	3.3	28.1	3.4	27.0	0.0	27.7†	1.1	21.7†	1.0	**25.4**	0.7	**26.3**†	0.6	**23.5**	0.5
	2K	38.8	0.3	37.2	2.8	27.3	2.8	**29.0**	0.0	**28.9**†	0.2	**22.2**†	0.2	23.2	0.0	24.5	0.5	21.9	0.5

Table 13: Performance of keyphrase generation models on the `nlp`, `astro` and `paleo` domains for present and absent keyphrases separately. Values in **bold** indicate best scores and † indicates significance over BART-FT.