High-Precision Extraction of Emerging Concepts from Scientific Literature

Daniel King[1], Doug Downey[1], Daniel S. Weld[1,2]
Allen Institute for AI, Seattle, WA[1]
Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA[2]
daniel,dougd,danw@allenai.org

ABSTRACT

Identification of new concepts in scientific literature can help power faceted search, scientific trend analysis, knowledge-base construction, and more, but current methods are lacking. Manual identification can't keep up with the torrent of new publications, while the precision of existing automatic techniques is too low for many applications. We present an unsupervised concept extraction method for scientific literature that achieves much higher precision than previous work. Our approach relies on a simple but novel intuition: each scientific concept is likely to be introduced or popularized by a single paper that is disproportionately cited by subsequent papers mentioning the concept. From a corpus of computer science papers on arXiv, we find that our method achieves a Precision@1000 of 99%, compared to 86% for prior work, and a substantially better precision-yield trade-off across the top 15,000 extractions. To stimulate research in this area, we release our code and data.[1]

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Information systems** → *Content analysis and feature selection.*

KEYWORDS

Concept extraction; scientific literature; citation graph

ACM Reference Format:
Daniel King[1], Doug Downey[1], Daniel S. Weld[1,2]. 2020. High-Precision Extraction of Emerging Concepts from Scientific Literature. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401235

1 INTRODUCTION

Scientific Concept Extraction (SCE) aims to automatically extract concepts discussed in the scientific literature. For example, given a corpus of information retrieval papers, we would like to extract *Attentive Collaborative Filtering* [3] and *Self-Taught Hashing* [21] as notable scientific concepts. Automatic SCE is necessary because, while some scientific terms are already catalogued in existing knowledge bases such as Wikipedia[2], the vast majority are not, due to the breadth and rapid progress of science. Three examples from the top concepts output by our method are *ELMo* [17], *gradient penalty* [7], and *asynchronous advantage actor-critic (A3C)* [15], which do not have Wikipedia pages, despite being introduced by papers that now have thousands of citations. Existing methods struggle with the challenge of distinguishing phrases that are simply *associated* with a concept (e.g. *popular conjecture* or *input graph*) from phrases that truly *are* a concept (e.g. *3-SUM Conjecture* or *graph coloring*). Accurate, high-coverage SCE could power many applications, and is a first step toward automatically constructing a comprehensive scientific knowledge base.

We propose a new method for SCE, ForeCite, based on the intuition that scientific concepts tend to be introduced or popularized by a single paper—one which is disproportionately cited by other papers. We encode this intuition in a simple unsupervised algorithm that ranks extracted phrases by how well they follow this citation pattern. In experiments with recent papers from the computer science domain, we find that ForeCite outperforms the CNLC [19] and LoOR [10] methods from previous work, achieving a better Precision-Yield curve across the top 15,000 extractions and much higher precision among the highest-scoring extractions.

In summary, we make the following contributions:

(1) We introduce ForeCite, a simple, unsupervised method for extracting high-precision conceptual phrases from the scientific literature,
(2) We perform quantitative and qualitative evaluation against other SCE methods, showing ForeCite outperforms existing work, improving Precision@1000 from 86% to 99%, and
(3) We release the dataset, code, and evaluation annotations used in our experiments.

2 TASK DEFINITION AND BACKGROUND

SCE is the task of extracting phrases that are scientific concepts from a corpus of academic papers. Precisely defining *concept* is difficult. In this paper, we define a concept as a phrase that could reasonably have an encyclopedic article (akin to a Wikipedia page) that would be of interest to multiple scientists. While subjective, this definition was sufficient to achieve the high end of moderate inter-annotator agreement (Section 4.2). By our definition, many phrases are obviously correct (e.g. *BERT* [4] or *deep learning*) or too ambiguous or vague (e.g. *multiple styles* or *deterministic mechanisms*), but there are also many less clear phrases (e.g. *relationship detection* or *shape analysis*). Phrases can be on the fence because they might be too general, or too specific. For example, *sentence* is

[1]https://github.com/allenai/ForeCite

[2]www.wikipedia.org

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401235

too general, and *speaker identification performance* too specific. See our GitHub repository for the instructions used by evaluators in our experiments.[3]

Two separate areas of work focus on tasks related to, but distinct from, SCE. Keyphrase extraction (see [16] for a recent survey) is the task of extracting important topical phrases at the document level. In contrast, we extract important phrases at the corpus level—and obtain higher precision and more specific phrases compared to reported results on keyphrase techniques. Our task also differs from *topic modeling* performed by e.g. Latent Dirichlet Allocation (LDA) [2]. In LDA, topics are distributions over the full vocabulary, while our extracted concepts are individual phrases.

2.1 Existing Methods

While a variety of phrase mining approaches are applicable to SCE [2, 5, 11, 13, 14, 18], the closest prior works to ours are two approaches [10, 19] that use the *term citation graph*, an important building block of our approach. The term citation graph is the subgraph of the full citation graph that includes only papers containing a specific term (e.g. the term citation graph for *neural networks* includes exactly the papers that mention *neural networks*, along with all of their citation edges).

Both previous approaches use the intuition that a term citation graph for a concept should be more dense than that of a non-concept. For example, a paper that mentions the concept *LSTM* is very likely to cite other papers that also mention *LSTM*. This results in a term citation graph for *LSTM* that is more dense than that of a random term. We describe the two previous approaches below. All methods rank a set of phrases by scoring each phrase independently.

The first method we compare with, *LoOR*, uses log-odds [10]:

$$\text{LoOR}(G_t) = \log(P(O(G_t)|\text{concept})) - \log(P(O(G_t)|\text{not concept}))$$

The LoOR score for a term citation graph G_t is the log probability of making the observation $O(G_t)$ given that G_t is a concept, minus the log probability of making the observation $O(G_t)$ given that G_t is not a concept. See [10] for details.

The second method we compare with, *CNLC*, builds upon LoOR, but is a simpler formula and normalizes for the size of the term citation graph [19]:

$$\text{CNLC}(G_t) = \frac{c_t}{n_t} - \frac{c}{N}$$

The CNLC score for a term citation graph G_t is the number of citation edges within G_t, c_t, divided by the number of papers in G_t, n_t, minus the number of citation edges from G_t to the rest of the corpus, c, divided by the number of papers in the full corpus, N. See [19] for details.

Both methods use different text preprocessing and datasets to generate the set of candidate phrases. Our preprocessing and dataset are detailed in Section 4.1.

[3]https://github.com/allenai/ForeCite/EVALUATION_INSTRUCTIONS.md

3 FORECITE

ForeCite is based on a different hypothesis about how the term citation graph tends to be structured for a scientific concept. Specifically, ForeCite assumes that concepts tend to be introduced or popularized by a *central* paper; and that other papers discussing the concept cite the central paper. We show that the term citation graph structure resulting from a central paper is a higher-precision signal than the graph density based signal used in prior work. Specifically, the ForeCite concept score:

$$\text{ForeCite}(G_t) = max_{p \in G_t} \log(f_t^p + 1) \cdot \frac{f_t^p}{f_t}$$

This score is a maximum over papers in a given term citation graph G_t, where each paper p is scored based on the number f_t^p of *future* papers that contain the term t *and* cite p, and f_t the total number of future papers that contain t. The intuition is that more citations within G_t to p is better (the log term), and a higher fraction of papers containing t that cite p is better (the ratio term). There are two additional details to the algorithm: (1) only papers with at least 3 citations are scored, and (2) we sample 500 of the future papers with the phrase in order to compute the ratio.

4 EXPERIMENTAL RESULTS

We now present experiments measuring precision and yield of ForeCite on SCE, in comparison to methods from prior work.

4.1 Dataset

Our corpus contains all arXiv[4] papers from 1999 to September 2019 in CS.* and stat.ML categories. We obtain the title, abstract, body text, and citations of the papers from the Semantic Scholar corpus [1], which uses ScienceParse[5] to perform extraction. Our corpus contains ~203,000 papers. We extract lemmatized noun phrases using spaCy [9] `en_core_web_md`, and normalize by removing stopwords using NLTK [12] English stopwords plus the word "using."

Our candidate phrases include all noun phrases that occur in any title from 1999-2018, resulting in ~293,000 candidate phrases from ~173,000 papers. We use citation information from papers outside this range, but candidate phrases must occur in at least one title within this range. Candidates are drawn from titles to increase efficiency of our experiments. Restricting candidates to noun phrases that appear in titles does limit the yield of all algorithms, but based on informal experiments with candidates from abstracts, we believe precision would be similar with other candidate sets.[6]

4.2 Human Evaluation

Evaluating SCE requires an expert and is labor-intensive. Additionally, our methods consider hundreds of thousands of candidate phrases, and we want to focus our evaluation on the fraction of phrases that are highly ranked by the methods, because high precision is required for many applications. Annotating a static gold set that is independent of the systems' outputs and is still large enough

[4]www.arxiv.org

[5]https://github.com/allenai/science-parse and https://github.com/allenai/spv2

[6]With candidates from abstracts, we measured p@10000 of 0.92, similar to the 0.93 for title candidates (Table 1).

Table 1: Precision@K for each method for K ∈ {100, 1000, 10000}, estimated using a sample of size 100. ForeCite achieves higher precision than the other methods for all K. Results in boldface indicate significantly greater performance ($p < 0.05$) than both baselines, computed using the Fisher Exact Test.[8]

	ForeCite	CNLC	LoOR
Precision@100	**1**	0.93	0.91
Precision@1000	**0.99**	0.86	0.81
Precision@10000	0.93	0.88	0.84

to explore the high-precision regime is intractable. Given these difficulties, we evaluated highly-ranked phrases from each method's output, and all evaluation was performed by the first author of this paper. We compute inter-annotator agreement with the second author of this paper, first calibrating on a sample of 20 extractions, and then computing agreement on a sample of 100 (both samples are drawn from the annotations used for evaluation in Figure 1). We achieve raw agreement of 88% and a Cohen's Kappa of 0.58. This falls into the high end of moderate agreement, which is reasonable given the subjectivity of the task. The disagreements impacted each method fairly equally, and were primarily due to the second evaluator being less generous regarding what qualifies as a concept. Additionally, none of the disagreements fall in the top 5,000 results of our method, reinforcing confidence in the high precision of the top concepts from our method. We also release all annotations used for evaluation. The disagreements are illustrative of the subjectivity of the task, so we list them here: *road scene*, *downlink*, *stabilizer*, *quasi-polynomial hitting set*, *human demonstration*, *full-diversity*, *local geometric features*, *unit quaternion*, *recurrent model*, *random formula*, *quadratic-time hardness*, and *retail performance*.[7] For future work, we would like to further validate our judgements in a user-facing application.

4.3 Experiments

As noted above, objective evaluation of SCE is difficult, and previous work has used a variety of different evaluation procedures. Jo et al. [10] evaluated by inspection, and ul Haque and Ginsparg [19] compared concept phrases against librarian-assigned keywords and searches on arXiv. Here, we evaluate the output of each method and measure precision. A measure of recall is not possible without exhaustive gold data, so we focus on the quality of the highly-ranked phrases in terms of Precision@K and Precision-Yield curves.

First, we present precision measurements at ranks of 100, 1000, and 10000 in Table 1. We evaluated a random sample of size 100 from the top-K ranked phrases of each method. Importantly, even at K of 100, the competing methods do not achieve a precision of 1. Our method does, and maintains precision near 1 out to K of 1,000.

Second, we evaluated a separate random sample of size 300 from the union of the top 15,000 phrases from each method. In Figure 1, we present a Precision-Yield curve for each method from these annotations. Each point on this curve corresponds to one positive annotation, and has an x value of the estimated true positive yield and a y value of the cumulative precision. Our method outperforms the baselines both at the high-precision end of the curve, and overall, resulting in a 38% reduction in area over the curve relative to CNLC and a 60% reduction relative to LoOR.

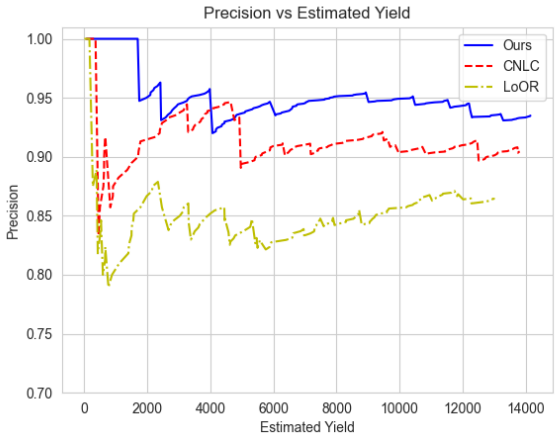

Figure 1: Precision-Yield curves of top 15,000 extractions. ForeCite outperforms the other methods.

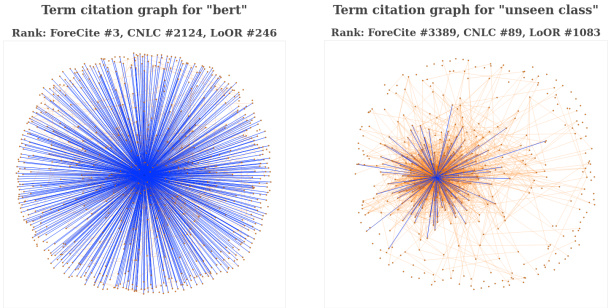

Figure 2: Illustrative term citation graphs for a concept highly ranked by ForeCite (*BERT*, left), and CNLC (*unseen class*, right) respectively. Blue edges connect to the central paper according to our method, other edges are orange. The *BERT* graph has a much higher proportion of links to the central paper than the *unseen class* graph.

5 DISCUSSION

5.1 Qualitative Method Comparison

Qualitatively, the different SCE methods favor different types of term citation graphs. We illustrate this difference in Figure 2, which plots the term citation graphs for two phrases, one highly ranked by ForeCite, one highly ranked by CNLC. ForeCite values a "spiky" graph structure with many links to a central paper, while CNLC and LoOR value dense graph structures indicative of a phrase shared amongst a citation community. This difference can be seen in Figure 2, where the *BERT* citation graph is dominated by links to

[7] The phrases are difficult to assess in isolation, but we encourage the reader to search for papers mentioning these phrases to see their use in context.
[8] https://www.socscistatistics.com/tests/fisher/default2.aspx are bolded

the central paper, colored blue (603 out of 678 nodes link to the max degree node), whereas the *unseen class* citation graph has a much higher proportion of links between other papers, colored orange (113 out of 415 nodes link to the max degree node). Additionally, LoOR does not normalize for the size of the term citation graph, so it favors phrases that occur very frequently in the corpus. The above differences between the methods suggest that ForeCite could be a helpful addition to existing tools.

The intuition behind ForeCite also leads it to produce more specific concepts. For example, the top five phrases ranked by ForeCite in our experiments are *fast gradient sign method*, *DeepWalk*, *BERT*, *node2vec*, and *region proposal network*. By comparison, CNLC's top five are *VQA*, *adversarial example*, *adversarial perturbation*, *ImageNet*, and *person re-identification*, and LoOR's top five are *codeword*, *received signal*, *achievable rate*, *convolutional layer*, and *antenna*. Given our goal of augmenting existing knowledge bases with new, specific, concept pages, we would like to know if we are extracting *emerging* concepts relative to Wikipedia. Wikipedia has specific inclusion criteria[9], but more recent and specific concepts are less likely to have Wikipedia pages, and these are the type of concepts that ForeCite tends to rank highly. As an indication of this, we measure how many of the top-20 phrases from each method already have Wikipedia pages, finding that only 30% of them do for our method, compared to 50% and 90% for CNLC and LoOR respectively.

Due to the high precision of ForeCite in the regimes measured in our experiments, our data includes only 18 unique errors for the method. An error analysis revealed that ten of the 18 phrases were too general or vague, whereas the other eight were too specific. Further, five errors were due to mistakes in PDF parsing or noun phrase extraction, rather than ForeCite's ranking heuristic.

5.2 Analysis of Central Papers

The intuition behind ForeCite is that valid concepts are generally introduced in a central paper, and ForeCite identifies this central paper. As verification of the importance of this intuition, we evaluated our top-100 concepts for whether the central paper does in fact introduce the concept, and found that it does for 95 of them. For example, the concept *fast gradient sign method* is associated with Goodfellow et al. [6], which introduces the *fast gradient sign method* as a way to generate adversarial examples. The content and citations of the introducing paper are a rich data source for downstream applications such as constructing a knowledge base entry for the concept; exploring this is an item of future work.

Applied to the two most cited papers from SIGIR 2017 (that are on arXiv) according to Microsoft Academic[10], our method correctly identifies the highest-scoring concept for each paper as *IRGAN* [20] and *Neural Factorization Machine* [8].

6 CONCLUSION AND FUTURE WORK

In this paper, we present a simple, unsupervised method for high-precision concept extraction from the scientific literature. We show that our method outperforms prior work using term citation graphs, particularly in the high-precision regime. In future work, we would like to apply our method to a corpus beyond arXiv computer science, use the output of our method as distant supervision for more powerful textual concept extraction, and use our concept extraction as a starting point for further applications, including semi-automated construction of encyclopedia pages for science.

[9]https://en.wikipedia.org/wiki/Wikipedia:Notability
[10]www.academic.microsoft.com

ACKNOWLEDGMENTS

This work was supported in part by NSF Convergence Accelerator award 1936940, ONR grant N00014-18-1-2193, and the University of Washington WRF/Cable Professorship. We thank Iz Beltagy and Oren Etzioni for helpful discussions, Madeleine van Zuylen for assistance with early evaluations, Marti Hearst for comments on a draft, and the rest of the Semantic Scholar team for comments on drafts, and assistance with the text processing and data collection.

REFERENCES

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Michael Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler C. Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna L. Power, Sam Skjonsberg, Lucy Lu Wang, Christopher Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *NAACL-HLT*.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wenjin Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR '17*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[5] Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting Keyphrases from Research Papers Using Citation Networks. In *AAAI*.

[6] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*. http://arxiv.org/abs/1412.6572

[7] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *NIPS*.

[8] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *ACM SIGIR 2017*. 355–364.

[9] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[10] Yookyung Jo, Carl Lagoze, and C. Lee Giles. 2007. Detecting research topics via the correlation between graphs and texts. In *KDD '07*.

[11] Jon M. Kleinberg. 2002. Bursty and hierarchical structure in streams. In *KDD*.

[12] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002).

[13] Sepideh Mesbah, Christoph Lofi, Manuel Valle Torre, Alessandro Bozzon, and Geert-Jan Houben. 2018. TSE-NER: An Iterative Approach for Long-Tail Entity Extraction in Scientific Publications. In *International Semantic Web Conference*.

[14] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order Into Texts. In *EMNLP 2004*.

[15] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *ICML*.

[16] Eirini Papagiannopoulou and Grigorios Tsoumakas. 2020. A review of keyphrase extraction. *WIREs Data Mining and Knowledge Discovery* 10, 2 (2020), e1339. https://doi.org/10.1002/widm.1339 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1339

[17] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*, Vol. 1. 2227–2237.

[18] Kumar Shubankar, Adityapratap Singh, and Vikram Pudi. 2011. A frequent keyword-set based algorithm for topic modeling and clustering of research papers. *2011 3rd Conference on Data Mining and Optimization (DMO)* (2011), 96–102.

[19] Asif ul Haque and Paul Ginsparg. 2011. Phrases as subtopical concepts in scholarly text. In *JCDL '11*.

[20] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *ACM SIGIR 2017*. 515–524.

[21] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. 2010. Self-taught hashing for fast similarity search. In *SIGIR '10*.