OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services

Xiao Liu*
Tsinghua University
liuxiao21@mails.tsinghua.edu.cn

Da Yin*
Tsinghua University
yd18@mails.tsinghua.edu.cn

Jingnan Zheng
National University of Singapore
e0718957@u.nus.edu

Xingjian Zhang
Tsinghua University
zhangxj18@mails.tsinghua.edu.cn

Peng Zhang
Zhipu AI
zpjumper@gmail.com

Hongxia Yang
DAMO Academy, Alibaba Group
yang.yhx@alibaba-inc.com

Yuxiao Dong
Tsinghua University
yuxiaod@tsinghua.edu.cn

Jie Tang†
Tsinghua University
jietang@tsinghua.edu.cn

ABSTRACT

Academic knowledge services have substantially facilitated the development of the science enterprise by providing a plenitude of efficient research tools. However, many applications highly depend on ad-hoc models and expensive human labeling to understand scientific contents, hindering deployments into real products. To build a unified backbone language model for different knowledge-intensive academic applications, we pre-train an academic language model OAG-BERT that integrates both the heterogeneous entity knowledge and scientific corpora in the Open Academic Graph (OAG)—the largest public academic graph to date. In OAG-BERT, we develop strategies for pre-training text and entity data along with zero-shot inference techniques. OAG-BERT achieves outperformance over baselines on nine academic tasks including two demo applications, demonstrating its potential to serve as one foundation model for academic knowledge services. Its zero-shot capability furthers the path to mitigate the need of expensive annotations. OAG-BERT has been deployed for real-world applications, such as the reviewer recommendation function for National Nature Science Foundation of China (NSFC)—one of the largest funding agencies in China—and paper tagging in AMiner (https://www.aminer.cn). All codes and pre-trained models are available via the CogDL toolkit[1].

CCS CONCEPTS

• **Information systems → Language models**; **Data mining**; • **Computing methodologies → Knowledge representation and reasoning**; **Supervised learning by classification**.

*The authors contributed equally to this research.
†Jie Tang is the corresponding author.
[1]https://github.com/thudm/oag-bert.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539210

KEYWORDS

Pre-Training; Language Model; Heterogeneous Knowledge Graph

ACM Reference Format:
Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022. OAG-BERT: Towards A Unified Backbone Language Model For Academic Knowledge Services . In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539210

1 INTRODUCTION

Academic knowledge services, such as AMiner [38], Google Scholar, Microsoft Academic Service [41], and Semantic Scholar, have been of great assistance to advance the science enterprise. Beyond collecting statistics, e.g., citation count, an increasing attention of these platforms has been focused on providing AI-powered academic knowledge applications, including paper recommendation [8, 14], expert matching [29], taxonomy construction [33], and knowledge evolution [48].

However, most of these applications are built with specified models to understand scientific contents. For example, OAG Zhang et al. employs the doc2vec [19] embeddings trained on a small corpus for academic entity alignment [49]. Zhang et al. [53] leverage an attention strategy to model the text and metadata embeddings for paper tagging. In other words, the success of such academic systems heavily rely on different language understanding components. In addition, task-specific annotated datasets required by these components demand arduously expensive labeling cost.

The newly emerging pre-trained models, such as BERT [10] and GPT [30], have substantially promoted the development of natural language processing (NLP). Specifically, pre-trained language models for academic data have also been developed, such as BioBERT [20] for the biomedical field and SciBERT [3] for scientific literature. However, these models mainly focus on scientific texts and ignore the connected entity knowledge that can be crucial for many knowledge-intensive applications. For example, in author name disambiguation [7, 54], the affiliations of a paper's authors offer important signals about their identities.

In light of these issues, we propose to pre-train a unified entity-augmented academic language model, OAG-BERT, as the backbone

model for diverse academic mining tasks and knowledge applications. OAG-BERT is pre-trained from the Open Academic Graph (OAG) [49], which is to date the largest publicly available heterogeneous academic entity graph. It contains more than 700 million entities (papers, authors, fields of study, venues, and affiliations), 2 billion relationships, and 5 million papers with full contents and 110 million abstracts as corpora.

To handle the heterogeneous knowledge, we design the entity type embedding for each type of entities, respectively. To implement the masked language pre-training over entity names with various lengths, we leverage a span-aware entity masking strategy that can select to mask a continuous span of tokens according to the entity length. To better "notify" the OAG-BERT model with the entity span and sequence order, we propose the entity-aware 2D positional encoding to take both the inter-entity sequence order and intra-entity token order into consideration.

We apply OAG-BERT to nine academic knowledge applications, including name disambiguation [7, 54], literature retrieval, entity graph completion [11, 16], paper recommendation [8], user activity prediction [8], fields-of-study tagging [26], venue prediction, affiliation prediction, and automatic title generation. Moreover, we present a number of prompt-based zero-shot usages of OAG-BERT, including the predictions of a paper's venue, affiliations, and fields of study, in which the annotation cost is significantly mitigated.

To sum up, we make the following contributions in this paper:

- **A Unified Backbone Model OAG-BERT:** We identify the challenge in existing academic knowledge applications, which heavily depend on ad-hoc models, corpora, and task-specific annotations. To address the problem, we present OAG-BERT as a unified backbone model with 110M parameters to support it.
- **Entity-Augmented Language Model Pre-Training:** In OAG-BERT, we enrich the language model with the massive heterogeneous entity knowledge from OAG. We design pre-training strategies to incorporate entity knowledge into the model.
- **Prompt-based Zero-Shot Inference:** We design a decoding strategy to allow OAG-BERT to perform well on prompt-based zero-shot inference, which offers the potential to significantly reduce the annotation cost in many downstream applications.
- **System Deployment and Open-Sourced Model:** We demonstrate the effectiveness of OAG-BERT on nine academic knowledge applications. In addition, OAG-BERT has been deployed as the infrastructure of AMiner[2] and also used for NSFC's grant reviewer recommendation. The pre-trained model is open to public access through the CogDL [6] package for free.

2 RELATED WORKS

Our proposed OAG-BERT model is based on BERT [10], a self-supervised [22] bidirectional language model. It employs multi-layer transformers as its encoder and uses masked token prediction as its objective, allowing massive unlabeled text data as a training corpus. BERT has many variants. SpanBERT [17] develops span-level masking which benefits span selection tasks. ERNIE [55] introduces explicit knowledge graph inputs to the BERT encoder and achieves significant improvements over knowledge-driven tasks.

[2]https://www.aminer.cn/

As for the academic domain, previous works such as BioBERT [20] or SciBERT [3] leverage the pre-training process on scientific domain corpus and achieve state-of-the-art performance on several academic NLP tasks. The S2ORC-BERT [25], applies the same method with SciBERT on a larger scientific corpus and slightly improves the performance on downstream tasks. Later works [15] further show that continuous training on specific domain corpus also benefits the downstream tasks. These academic pre-training models rely on large scientific corpora. SciBERT uses the semantic scholar corpus [2]. Other large academic corpora including AMiner [38], OAG [38, 49], and Microsoft Academic Graph (MAG) [18] also integrate massive publications with rich graph information as well, such as authors and research fields.

On academic graphs, some tasks involve not only text information from papers but also structural knowledge lying behind graph links. For example, to disambiguate authors with the same names [7, 54], the model needs to learn node representations in the heterogeneous graph. To better recommend papers for online academic search [13, 14], graph information including related academic concepts and published venues could provide great benefits. To infer experts' trajectory across the world [44], associating authors with their affiliation on semantic level would help. Capturing features from paper titles or abstracts is far from enough for these types of challenges.

3 OAG-BERT: A LANGUAGE MODEL WITH ACADEMIC KNOWLEDGE

The proposed OAG-BERT model is a bidirectional Transformer-based pre-training model by following conventional BERT architecture with 12 transformer [40] encoder layers. To provide unified support for various academic knowledge data mining applications, despite existing models like SciBERT [3] pre-trained over academic corpus, they ignore to incorporate enough academic entity knowledge. Based on OAG, the world's largest public academic heterogeneous entity graphs together with academic corpus, we propose to enrich OAG-BERT with OAG's entity knowledge.

However, this ambition also brings new challenges, as the original BERT architecture only focuses on natural language pre-training. To accomplish the goal, we propose several novel improvements to the model architecture and the pre-training process to allow efficient grasping of knowledge. We will introduce them in the following sections.

3.1 Model Architecture

The key challenge for OAG-BERT is how to integrate knowledge into language models. Previous approaches [21, 55] mainly focus on injecting homogeneous entities and relations from knowledge graphs like Wikidata, and few study heterogeneous entities.

To augment OAG-BERT with various types of entity knowledge, we place title, abstract and other entities from the same paper in a single sequence as one training instance. Figure 1 illustrates one example. There are five types of entities in total. We treat the text features (title and abstract) of a paper as one special text entity. The published venue, authors, affiliations, and research fields are the other four types of entities. Following the notation in OAG [49], we use fields-of-study (FOS) to denote research fields. Thanks to

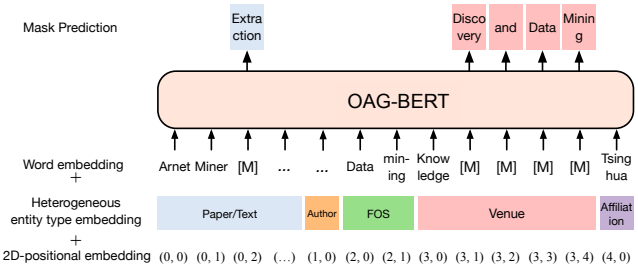

Figure 1: Heterogeneous entity augmentation in OAG-BERT. 1) *Heterogeneous entity type embedding* allows OAG-BERT be aware of different types of entities, 2) *Span-aware entity masking* selects a continuous span within long entities (such as the "Knowledge Discovery and Data Mining"), and 3) *Entity-aware 2D-positional embedding* jointly models inter and intra-entity token orders.

OAG, the entity names have been cleaned up, deduplicated, and unified, which enables OAG-BERT to learn consistent entity representations.

All the entities from one paper are concatenated as an input sample. To help the OAG-BERT model distinguish them, we use another three techniques: *Heterogeneous entity type embedding*, *Entity-aware 2D-positional encoding*, and *Span-aware entity masking*.

Heterogeneous entity type embedding. In order to distinguish different types of entities, we propose to leverage entity type embedding in the pre-training process to indicate entity type, whose usage is similar to the token type embedding used in BERT.

For example, given the title and abstract of a paper "ArnetMiner: extraction and mining of academic social networks", we retrieve its authors, fields of studies, venues, and affiliation entities and concatenate them into a sequence less than 512 tokens. For pure text (such as title and abstract), we label them with the original entity type index (e.g., 0) to acquire its entity type embedding. For author entities (such as "Jie Tang"), we label them with author type index (e.g., 1). So are for other entities. What's more, because entities should be order-invariant in the sequences, we shuffle their order in a sample sequence to avoid our model to learn any positional biases of these entities.

Entity-aware 2D-positional encoding. It is known that the transformer [40] is permutation-invariant (i.e. unaware of the sequence order) and the critical technique to indicate the sequence order in languages is to add *positional embedding*. However, to capture entity knowledge, the existing positional embeddings for pure texts are generally not applicable, as they cannot distinguish words from entities adjacent to each other and of the same type. For instance, if there are two affiliations "Tsinghua University" and "University of California" being placed next to each other in a sequence, the transformer would assume that there is an affiliation named "Tsinghua University University of California".

To sum up, our requests could be summarized in two points: 1) the positional embedding should imply the *inter-entity* sequence order (which is used to distinguish different entities) and 2) the positional embedding should indicate the *intra-entity* token sequence order (which is used as the traditional positional embedding).

In light of this, we design the entity-aware 2D-positional embedding that solves both the inter-entity and intra-entity problem (Cf. Figure 1). The first dimension is for inter-entity order, indicating which entity the token is in; the second dimension is for intra-entity order, indicating the sequence of tokens. For a given position, the final positional embedding is calculated by adding the two positional embeddings together.

Span-aware entity masking. When performing masking, for pure text contents such as paper titles and abstracts, we adopt the same random masking strategy as in BERT. But for heterogeneous academic entities, we expect OAG-BERT to memorize them well and thus develop a span-aware entity masking strategy combining the advantages of both ERNIE [55] and SpanBERT [17].

The intuition of using this strategy is that, some of the entities are too long for OAG-BERT to learn when using random masking at single-token granularity. Our span-aware entity masking strategy not only alleviates the problem, but also preserves the sequential relationship of an entity's tokens: for an entity that has less than 4 tokens, we will mask the whole entity; and for others, we sample masked lengths from a geometric distribution $\text{Geo}(p)$ which satisfies:

$$p = 0.2\,, \text{and } 4 \leq \text{Geo(p)} \leq 10 \tag{1}$$

If the sampled length is less than the entity length, we will only mask out the entity. For text contents and entity contents, we mask 15% of the tokens for each respectively.

Pre-LN BERT. Except for the previous changes to the original BERT architecture, we further adopt the Pre-LN BERT as used in deepspeed [31] , where layer normalization is placed inside the residual connection instead of after the add-operation in Transformer blocks. Previous work [52] demonstrates that training with Pre-LN BERT avoids vanishing gradients when using aggressive learning rates. Therefore, it is more stable than the traditional Post-LN version for optimization.

3.2 Implementation Details

The pre-training of OAG-BERT is separated into two stages. In the first stage, we only use scientific texts (paper title, abstract, and body) as the model inputs, without using the entity augmented inputs introduced above. This process is similar to the pre-training of the original BERT model. We name the intermediate pre-trained model as the vanilla version of OAG-BERT. In the second stage, based on the vanilla OAG-BERT, we continue to train the model on the heterogeneous entities, including titles, abstracts, venues, authors, affiliations, and FOS.

First Stage: Pre-train the vanilla OAG-BERT. In the first stage of pre-training, we construct the training corpus from two sources: one comes from the PDF storage of AMiner; and the other comes from the PubMed XML dump. We clean up and sentencize the corpus with SciSpacy [27]. The corpus adds up to around 5 million unique paper full-text from multiple disciplines. In terms of vocabulary, we construct our OAG-BERT vocabulary using WordPiece, which is also used in the original BERT implementation. This ends up with 44,000 unique tokens in our vocabulary. The original BERT employs a sentence-level loss, namely Next Sentence Prediction

(NSP), to learn the entailment between sequences, which has been found not that useful [17, 24] and thus we do not adopt it.

To better handle the entity knowledge of authors in the OAG, we transform the author name list into a sentence for each paper and place it between the title and abstract in the data prepossessing. Therefore, compared to previous models like SciBERT, our vocabulary contains more tokens from author names. Following the training procedures of BERT, the vanilla OAG-BERT is first pre-trained on samples with a maximum of 128 tokens and then shift to pre-training it over samples with 512 tokens.

Second Stage: Enrich OAG-BERT with entity knowledge. In the second stage of pre-training, we use papers and related entities from the OAG corpus. Compared to the corpus used in the first stage, we do not have full texts for all papers in OAG. Thus, we only use paper title and abstract as the paper text information. From this corpus, we select all authors with at least 3 papers published. Then we filter out all papers not linked to these selected authors. Finally, we got 120 million papers, 10 million authors, 670 thousand FOS, 53 thousand venues, and 26 thousand affiliations. Each paper and its connected entities are concatenated into a single training instance, following the input construction method described above. In this stage, we integrate the three strategies mentioned in Section 3.1 to endow OAG-BERT the ability to "notice" the entities, rather than regarding them as pure texts. For tasks that require document-level representations, we present a version of OAG-BERT with additional task-agnostic triplet contrast pre-training, which uses papers from the same authors in OAG as the positive pair and papers from authors with similar names as the negative pair.

Our pre-training is conducted with 32 Nvidia Tesla V100 GPUs and an accumulated batch size of 32768. We use the default BERT pre-training configurations in deepspeed. We run 16K steps for the first stage pre-training and another 4K steps for the second stage.

4 APPLICATIONS

We choose 9 fundamental academic mining tasks that are either directly deployed in the AMiner system or serve as prerequisites to other academic knowledge services, in which the entity knowledge may play an indispensable role. These applications feature 5 typical downstream applications:

- Author Name Disambiguation [7, 37, 54]
- Scientific Literature Retrieval [8, 42]
- Paper Recommendation [13, 14, 36]
- User Activity Prediction [8, 47, 50]
- Entity Graph Completion [4, 12, 16]

and 4 prompt-based zero-shot applications without need for any annotations:

- Fields-of-study Tagging [26, 34]
- Venue Prediction [1, 46]
- Affiliation Prediction [43, 44]
- Automatic Title Generation [28, 51]

We take SciBERT [3] as our major compared method to demonstrate the importance of our entity-augmented pre-training. Other baselines compared are introduced individually in each section.

4.1 Downstream applications

Compared to language models for common NLP tasks, backbone language models for academic mining are usually combined with other downstream supervised learning algorithms, such as clustering for name disambiguation [7, 54] and graph neural networks for relation completion [16]. This requires our language model to provide more informative representations on different entities.

Conventionally, these applications rely on individually trained representation upon their own small dataset or corpus. For example, in [16], to acquire embeddings for heterogeneous entities such as venues, fields-of-study and affiliations, the authors leverage metapath2vec [11] embeddings which contains no semantic information; in [54] authors use word2vec embeddings trained on a small portion of paper abstracts from AMiner systems. As an effort to unifying infrastructure for these applications, in the following evaluation OAG-BERT reports to achieve better performance on all of them.

Table 1: The Macro Pairewise F1 scores for the author name disambiguation competition whoiswho-v1.

	Inputs	SciBERT	OAG-BERT
Unsupervised	*title*	0.3690	**0.4120**
	+fos	0.4101	**0.4643**
	+venue	0.3603	**0.4247**
	+fos+venue	0.3903	**0.4823**
Supervised	Leader Board Top1	0.4900	

Author Name Disambiguation. Name disambiguation, or namely "disambiguating who is who", is a fundamental challenge for curating academic publication and author information, as duplicated names widely exist in our lives. For example, Microsoft Academic reports more than 10,000 authors named "James Smith" in United States [35]. Without effective author name disambiguation algorithms, it is difficult to identify the belonging-ship of certain papers for supporting applications such as expert matching, citation counting and h-index computing.

Given a set of papers with authors of the same name, the problem is usually formulated as designing algorithm to separate these papers into clusters, where papers in the same cluster belong to the same author and different clusters represent different authors. We use the public dataset *whoiswho-v1* [7, 54][3] and apply the embeddings generated by pre-trained models to solve name disambiguation from scratch. Following dataset setting, for each paper, we use the paper title and other attributes such as FOS or venue as input. We average over all the output token embeddings for title as the paper embedding. Then, we build a graph with all papers as the graph nodes and set a threshold to select edges. The edges are between papers where the pairwise cosine similarity of their embeddings is larger than the threshold. Finally, for each connected component in the graph, we treat it as a cluster. We searched the thresholds from 0.65 to 0.95 on the validation set and calculated the macro pairwise f1 score on test.

The results in Table 1 indicate that the embedding of OAG-BERT is significantly better than the SciBERT embedding while directly used in the author name disambiguation. We also observe that for

[3] https://www.aminer.cn/whoiswho

SciBERT the best threshold is always 0.8 while this value for OAG-BERT is 0.9, which reflects that the paper embeddings produced by OAG-BERT are generally closer than the ones produced by SciBERT.

In Table 1 we list a range of experimental results given title, field-of-study, and venue as inputs respectively. Though we attempted to use the abstract, author, and affiliation information, there is no performance improvement as expected. We speculate it is because these types of information are more complex to use, which might require additional classifier head or fine-tuning, as the supervised classification task mentioned above. In addition, we also report the top 1 score in the name disambiguation challenge leaderboard[4] and find that our proposed OAG-BERT reaches close performance as compared with the top-1 ad-hoc model for the contest.

Table 2: Scientific Literature Retrieval evaluation on OAG-QA (Top-100) between SciBERT and OAG-BERT.

	SciBERT	OAG-BERT
Geometry	0.097	**0.147**
Math. & Stats.	0.099	**0.166**
Algebra	**0.071**	0.069
Calculus	0.091	**0.160**
Number theory	0.067	**0.085**
Linear algebra	0.111	**0.160**
Astrophysics	0.041	**0.072**
Quantum mechanics	0.047	**0.080**
Classical mechanics	0.085	**0.197**
Chemistry	0.181	**0.216**
Biochemistry	0.146	**0.319**
Health care	0.041	**0.262**
Natural science	0.101	**0.277**
Algorithm	0.084	**0.209**
Neuroscience	0.054	**0.120**
Computer vision	0.035	**0.205**
Data mining	0.082	**0.161**
Deep learning	0.044	**0.138**
Machine learning	0.085	**0.177**
NLP	0.05	**0.160**
Economics	0.055	**0.151**
Average	0.079	**0.168**

Scientific Literature Retrieval. Scientific literature retrieval, which assists researchers finding relevant scientific literature given their natural language queries, is closely related to a wide-range of top-level applications including publication search, citation prediction and scientific question and answering. For example, for a professional question like *"Does sleeping fewer hours than needed cause common cold?"*, we may retrieve a related paper *"Sick and tired: does sleep have a vital role in the immune system?"*.

We evaluate OAG-BERT with triplet contrastive training over a fine-grained topic-specific literature retrieval dataset OAG-QA, which is constructed by collecting high-quality pairs of questions and cited papers in answers from Online Question-and-Answers (Q&A) forums (Stack Exchange and Zhihu). It consists of 22,659 unique query-paper pairs from 21 scientific disciplines and 87 fine-grained topics. Given each topic is accompanied by 10,000 candidate papers including the groundtruth, and their titles and abstracts are taken as the corpus. We compute the cosine similarity between output embeddings of the query and paper for ranking. Results in Table 2 suggest that OAG-BERT has a consistently better performance than SciBERT across 20 scientific disciplines .

[4]https://www.biendata.xyz/competition/aminer2019/leaderboard/

Table 3: Paper recommendation and User Activity Prediction (Co-View and Co-Read) on Scidocs [8].

Models	Paper Rec.		Co-View		Co-Read	
	nDCG	P@1	MAP	nDCG	MAP	nDCG
Random	51.3	16.8	25.2	51.6	25.6	51.9
doc2vec	51.7	16.9	67.8	82.9	64.9	81.6
Sent-BERT	51.6	17.1	68.2	83.3	64.8	81.3
SciBERT	52.1	17.9	50.7	73.1	47.7	71.1
OAG-BERT	**52.6**	**18.6**	**74.7**	**86.3**	**71.4**	**84.7**

Paper Recommendation & User Activity Prediction. As the number of scientific publications keeps soaring up, paper recommendation is playing an increasingly crucial role in many online academic systems, and therefore it is important to evaluate a backbone model's ability in boosting a production recommendation system. We consider the situation when users are browsing certain papers in our systems, and we want to 1) recommend them related papers of the ones they are reading, 2) predict papers they simultaneously viewed (Co-View) or pdf-accessed (i.e., Co-Read) in a user's browser session. In practice, for paper recommendation, it is often conducted in an ensemble manner: together with cosine similarities of textual embeddings encoded by language models, we jointly take other features such as citation overlaps, clicking counts and author similarities into consideration, and train a classifier to make the final decision; for user activity prediction, we mainly measure co-viewed or co-read papers' textual similarities.

We adopt Scidocs [8] paper recommendation and user activity prediction dataset for offline evaluation, which is constructed from real user clickthroughs and loggings in a publication search engine. The recommendation dataset consists of 22k samples, in which 20k clickings are used for training the recommender, 1k for validation, and 1k for testing. For Co-View and Co-Read dataset, each of them contains 30k papers. Besides SciBERT, we compare with common passage representation methods, including doc2vec [19] and Sent-BERT [32]. Results in Table 3 show that OAG-BERT with triplet contrastive training can bring a consistent gain over compared methods in both paper recommendation and user activity prediction setting.

Entity Graph Completion. Academic entity graph, which consists of heterogeneous entities including papers, authors, fields-of-study, venues, affiliations and other potential entities with attributes, is a powerful organization form of academic knowledge and finds wide adoptions in many academic systems such as Microsoft Academic Graph [35] and AMiner [38]. However, such entity graphs have been suffered from the long-standing challenge of incomplete and missing relations, and therefore the task of entity graph completion becomes vital to their maintenance.

In this section, we apply the heterogeneous entity embeddings of OAG-BERT as pre-trained initialization for entity embeddings on the academic graph and show that OAG-BERT can also work together with other types of models. Specifically, we take the heterogeneous graph transformer (HGT) model from [16], a state-of-the-art graph neural network, to conduct entity graph completion pre-trained embeddings from OAG-BERT.

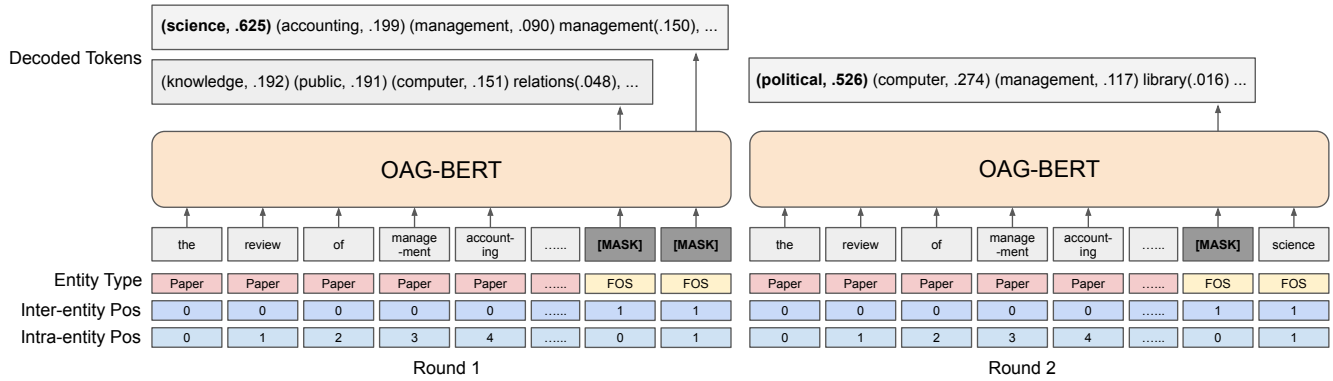

Figure 2: The decoding process of OAG-BERT. The left figure indicates that OAG-BERT decodes the masked token "science" at the second position with the highest probability (0.625) for the first round. Then it decodes "political" at the first position with the highest probability (0.526) for the second round as shown in the right figure.

Table 4: Results on Entity Graph Completion using HGT. OAG-BERT yields better initialization for heterogeneous entities.

Models	Paper-Field		Paper-Venue	
	NDCG	MRR	NDCG	MRR
XLNet	0.3939	0.4473	0.4385	0.2584
SciBERT	0.4740	0.5743	0.4570	0.2834
OAG-BERT	**0.4892**	**0.6099**	**0.4844**	**0.3131**

To make predictions for the links in the heterogeneous graph, the authors of HGT first extract node features and then apply HGT layers to encode graph features. For paper nodes, the authors use XLNet [45], a well-known general-domain pre-trained language model, to encode titles as input features. For other types of nodes, HGT use metapath2vec [11] to initialize the features. However, XLNet was pre-trained on universal language corpus, lacking academic domain knowledge, and can only encode paper nodes by using their titles and is unable to generate informative embeddings for other types of nodes.

To this end, we replace the original XLNet encoder with our OAG-BERT model, which can tackle the two challenges mentioned above. We use the OAG-BERT model to encode all types of nodes and use the generated embeddings as their node features. To demonstrate the effectiveness of OAG-BERT on encoding heterogeneous nodes, we also compare the performance of SciBERT with OAG-BERT. We experimented on the CS dataset released by HGT[5]. The details of the dataset are delivered in the appendix. The NDCG and MRR scores for the Paper-Field and Paper-Venue link prediction are reported in Table 4. It shows that SciBERT surpasses the original XLNet performance significantly, due to the pre-training on the large scientific corpus. Our proposed OAG-BERT made further improvements on top of that, as it can better understand the entity knowledge on the heterogeneous graph.

4.2 Prompt-based Zero-shot Applications

Despite OAG-BERT's qualification in providing unified support to various downstream applications to get rid of ad-hoc models and corpus, a more challenging topic is to reduce task-specific annotations, which can be expensive in business deployment.

[5]https://github.com/acbull/pyHGT

Table 5: The results for zero-shot inference tasks.

Method	Paper Tagging		Venue		Affiliation	
	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR
SciBERT	19.93%	0.37	9.87%	0.22	6.93%	0.19
+prompt	29.59%	0.47	10.03%	0.21	8.00%	0.20
+abstract	25.66%	0.43	18.00%	0.32	10.33%	0.22
+both	35.33%	0.52	9.83%	0.22	12.40%	0.25
OAG-BERT	34.36%	0.51	21.00%	0.37	11.03%	0.24
+prompt	37.33%	0.55	22.67%	0.39	11.77%	0.25
+abstract	**49.59%**	**0.67**	**39.00%**	**0.57**	**21.67%**	**0.38**
+both	49.51%	**0.67**	38.47%	**0.57**	21.53%	**0.38**

Take affiliation prediction as an example, a common approach is to train a k-class classifier for k given candidate institutions. However, as the progress of science, new universities, laboratories and companies emerge and to incorporate them into the pool may require re-annotating and re-training of the classifier with considerably high cost.

In light of the recent prompt-based [23] zero-shot and few-shot advances of large-scale pre-trained language models such as GPT-3 [5], in this section we also explore the potential of applying OAG-BERT to zero-shot applications in academic mining. We discover that OAG-BERT works surprisingly well on some fundamental applications, such as paper tagging, venue/affiliation prediction, and generation tasks such as title generation. We will first introduce how we implement the zero-shot inference on OAG-BERT, and then the details of our applications.

OAG-BERT's zero-shot inference strategies. Although not using a unidirectional decoder structure like GPT-3, we find that the bidirectional encoder-based OAG-BERT is also capable of decoding entities based on the knowledge it learned during pre-training. A running-example is provided in Figure 2. In MLM, the token prediction can be seen as maximizing the probability of masked input tokens, treating predictions on each token independently by maximizing $\sum_{w \in masked} \log P(w|C)$, where *masked* is the collection of masked tokens and C denotes contexts. But in entity decoding,

Figure 3: Deployed zero-shot paper tagging service in AMiner. OAG-BERT yields fine-grained tags at various lengths.

we cannot ignore the dependencies between tokens in each entity, and thus need to jointly consider the probability of all tokens in one entity as following $\log P(w_1, w_2, ..., w_l|C)$, where l is the entity length and w_i is the i-th token in the entity. As MLM is not unidirectional model, the decoding order for the tokens in one entity can be arbitrary. Suppose the decoding order is $w_{i_1}, w_{i_2}, ..., w_{i_l}$, where $i_1, i_2, ..., i_l$ is a permutation of $1, 2, ..., l$. Then the prediction target can be reformed as maximizing

$$\sum_{1 \le k \le l} \log P(w_{i_k}|C, w_{i_1}, w_{i_2}, ..., w_{i_{k-1}}) \quad (2)$$

As the solution space is getting larger as l increases, we adopt the following two strategies to determine the decoding order:

- **Greedy:** we use greedy selection to decide the decoding order, by choosing the token with maximal probability to decode. An example is depicted in Figure 2.
- **Beam search:** we can also use beam search [39] to search the token combinations with the highest probability.

Another challenge lies in choosing the appropriate entity length. Instead of using a fixed length, we traverse all entity lengths in a pre-defined range depending on the entity type and choose top candidates according to the calculated probability in Equation 2.

Fields-of-study Tagging. Fields-of-study tagging, referred to as fields-of-study (FOS) linking, is a fundamental mission to associate unstructured scientific contents with structured disciplinary taxonomy (a case is presented in Figure 3). Its results also serve as indispensable features for various downstream supervised applications.

However, it is a notoriously arduous undertaking to discover new FOS from enormous corpora; in addition, how to continuously drive algorithms to link a paper with newly discovered FOS also remains largely unexplored. However, thanks to the massive entity knowledge OAG-BERT has grasped in pre-training, it can be solved now using OAG-BERT's zero-shot inference without a lift of fingers.

We present a case study, where OAG-BERT is applied to tag the paper of GPT-3 [5] given its title and abstract. Using beam search with a width of 16 to decode FOS entities, we search from single-token entities to quadruple-token entities. The top 16 generated ones are listed in Table 6. The groud truth (or namely *gold*) FOS later annotated in MAG are all included in the top 16. Surprisingly,

Table 6: OAG-BERT's zero-shot paper tagging on the paper of GPT-3 given its title and abstract. The groundtruth FOS are **bolded**. Newly created FOS by OAG-BERT are underlined.

Title	Language Models are Few-Shot Learners
Abstract	Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally...
Generated FOS	Natural language processing, Autoregressive language model, **Computer science**, Sentence, Artificial intelligence, Domain adaptation, **Language model**, Few shot learning, Large corpus, Arithmetic, Machine learning, Architecture, Theoretical computer science, Data mining, **Linguistics**, Artificial language processing
Gold FOS	Language model, Computer science, Linguistics

some fine-grained correct entities, though not in existing FOS, are also generated, such as *Autoregressive language model* or *Few shot learning*. Despite some ill-formed or inappropriate entities such as *Architecture* or *Artificial language processing*, OAG-BERT's zero-shot tagging capability is still quite amazing.

To quantitatively evaluate the performance paper tagging, we adapt FOS prediction task from MAG. First, we choose 19 top-level field-of-studies (FOS) such as "biology" and "computer science". Then, from the paper data which were not used in the pre-training process, we randomly select 1,000 papers for each FOS. The task is to rank all FOS for each paper by estimating the probabilities of Equation 2 given paper title and optional abstract.

We also apply two techniques to improve the model decoding performance. The first technique is to add extra *prompt* word to the end of the paper title (before masked tokens). We select "Field of study:" as the prompt words in the FOS inference task. The second technique is to concatenate the paper abstract to the end of the paper title. We report the Hit@1 and MRR scores in Table 5.

Venue and Affiliation Prediction. Analogously to paper tagging, venue and affiliation prediction of certain papers can also be conducted in zero-shot learning setting. From non-pretrained papers, we choose the 30 most frequent arXiv categories and 30 affiliations as inference candidates, with 100 papers randomly selected for each candidate. Full lists of the candidates including FOS candidates are enclosed in the appendix.

The experiment settings completely follow the FOS inference task, except that we use "Journal or Venue:" and "Affiliations:" as prompt words respectively. The entity type embeddings for masked entities in OAG-BERT are also replaced by venue and affiliation entity type embeddings accordingly.

In Table 5, we can see that the proposed augmented OAG-BERT outperforms SciBERT by a large margin. Although SciBERT was not pre-trained with entity knowledge, it still performs much greater than a random guess, which means the inference tasks are not independent of the paper content information. We speculate that the pre-training process on paper content (as used in SciBERT) also helps the model learn some generalized knowledge on other types of information, such as field-of-studies or venue names.

We also observe that the proposed use of abstract can always help improve the performance. On the other hand, the prompt

Figure 4: Deployed zero-shot Title Generation application for Turing Test in Wudao.[7]

words works well with SciBERT but only provide limited help for OAG-BERT. Besides, the affiliation inference task appears to be harder than the other two tasks. Further analysis are provided in the A.1. Two extended experiments are enclosed as well, which reveal two findings:

(1) Using the summation of token log probabilities as the entity log probability is better than using the average.
(2) The out-of-order decoding is more suitable for encoder-based models like SciBERT and OAG-BERT, compared with the left-to-right decoding.

Automatic Title Generation. How to summarize the contributions of a research paper into one sentence? Given the abstract, even senior experts may not figure out in a few seconds, but OAG-BERT can generate titles comparable to original human-written ones in a zero-shot manner. During the pre-training,the span masking strategy will also be applied to titles, allowing OAG-BERT to learn to summarize. Some case studies are presented in Table 7, in which we observe that OAG-BERT can generate quite the same title as origin given the paper abstract, even for our paper itself.

We also provide an interactive testing demo application online (as shown in Figure 4)[6] to test if graduate-level students can distinguish between OAG-BERT generated and original titles. Results suggest that there is probably only a small gap in performance between OAG-BERT's generation and human assignment's.

5 DEPLOYED APPLICATIONS

In this section, we will introduce several real-world applications where our OAG-BERT is employed.

First, the results on the name disambiguation tasks indicate that the OAG-BERT is relatively strong at encoding paper information with multi-type entities, which further help produce representative embeddings for the paper authors. Thus, we apply the OAG-BERT to the NSFC reviewer recommendation problem [9]. The National Natural Science Foundation of China is one of the largest science foundations, where an enormous number of applications are reviewed every year. Finding appropriate reviewers for applications is time-consuming and laborious. To tackle this problem, we collaborate with Alibaba and develop a practical algorithm on top of the OAG-BERT which can automatically assign proper reviewers to applications and greatly benefits the reviewing process.

In addition to that, we also integrate the OAG-BERT as a fundamental component for the AMiner [38] system. In AMiner, we utilize OAG-BERT to handle rich information on the academic heterogeneous graph. For example, with the ability of decoding FOS entities, we use the OAG-BERT to automatically generate FOS candidates for unlabeled papers. Besides, we similarly amalgamate the OAG-BERT into the name disambiguation framework. Finally, we employ OAG-BERT to recommend related papers for users, leveraging its capability in encoding paper embeddings. The OAG-BERT model is also released in CogDL package.

Table 7: Upper: case study in OAG-BERT generated titles and original title. **Lower:** Online testing result from 660 random human views on 50 pairs of OAG-BERT generated and original titles.

	OAG-BERT Generated v.s. Original
OAG-BERT	OAG-LM: A Unified Backbone for Academic Knowledge Services OAG-LM: A Unified Backbone Language Model for Academic Knowledge Services
AMiner	ArnetMiner: A System for Extracting and Mining Academic Social Networks ArnetMiner: Extraction and Mining of Academic Social Networks
ResNet	Deep Residual Networks for Visual Recognition : A Comparison of Deep and VGG Networks Deep Residual Networks for Image Recognition
SciBERT	SciBERT: A Pretrained Language Model for Scientific NLP SciBERT: A Pretrained Language Model for Scientific Text

Method	Total	Select	Selection Rate
OAG-BERT Generated	330	157	47.6%
Original	330	163	52.4%

6 CONCLUSION

In conclusion, we propose OAG-BERT, a heterogeneous entity-augmented language model to serve as the backbone for academic knowledge services. It incorporates entity knowledge during pre-training, which benefits many downstream tasks involving strong entity knowledge. OAG-BERT is applied to 9 typical selected academic applications and being deployed in AMiner system and NSFC's reviewer recommendation process. We finally release the pre-trained model in CogDL, providing free use to arbitrary users.

ACKNOWLEDGEMENT

We thank the reviewers for their valuable feedback to improve this work. This work is supported by Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108402, Natural Science Foundation of China (Key Program, No. 61836013), and National Science Foundation for Distinguished Young Scholars (No. 61825602).

[6]Try demo at: https://wudao.aminer.cn/turing-test/v1/game/pubtitle

REFERENCES

[1] Hamed Alhoori and Richard Furuta. 2017. Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics* 11, 2 (2017), 553–563.

[2] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the Literature Graph in Semantic Scholar. In *NAACL*. 84–91.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP*.

[4] Nesserine Benchettara, Rushed Kanawati, and Celine Rouveirol. 2010. Supervised machine learning applied to link prediction in bipartite social networks. In *ASONAM*. IEEE, 326–330.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NIPS* (2020).

[6] Yukuo Cen, Zhenyu Hou, Yan Wang, Qibin Chen, Yizhen Luo, Xingcheng Yao, Aohan Zeng, Shiguang Guo, Peng Zhang, Guohao Dai, et al. 2021. CogDL: an extensive toolkit for deep learning on graphs. *arXiv preprint arXiv:2103.00959* (2021).

[7] Bo Chen, Jing Zhang, Jie Tang, Lingfan Cai, Zhaoyu Wang, Shu Zhao, Hong Chen, and Cuiping Li. 2020. CONNA: Addressing Name Disambiguation on The Fly. *TKDE* (2020).

[8] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*. 2270–2282.

[9] David Cyranoski. 2019. Artificial intelligence is selecting grant reviewers in China. *Nature* 569, 7756 (2019).

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

[11] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*.

[12] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V Chawla, Jinghai Rao, and Huanhuan Cao. 2012. Link prediction and recommendation across heterogeneous social networks. In *ICDM*. IEEE, 181–190.

[13] Zhengxiao Du, Jie Tang, and Yuhui Ding. 2018. Polar: Attention-based cnn for one-shot personalized article recommendation. In *ECML-PKDD*. Springer.

[14] Zhengxiao Du, Jie Tang, and Yuhui Ding. 2019. POLAR++: Active One-shot Personalized Article Recommendation. *TKDE* (2019).

[15] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964* (2020).

[16] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WWW*.

[17] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL* 8 (2020).

[18] Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. 2019. A scalable hybrid research paper recommender system for microsoft academic. In *WWW*.

[19] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. PMLR, 1188–1196.

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020).

[21] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*, Vol. 34.

[22] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[23] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385* (2021).

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[25] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).

[26] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *HT Conference*. 31–40.

[27] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669* (2019).

[28] Vahed Qazvinian and Dragomir Radev. 2008. Scientific Paper Summarization Using Citation Summary Networks. In *COLING*. 689–696.

[29] Yujie Qian, Jie Tang, and Kan Wu. 2018. Weakly learning to match experts in online community. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 3841–3847.

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019).

[31] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*.

[32] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*. 3982–3992.

[33] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *WWW*. 486–497.

[34] Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A Web-scale system for scientific knowledge exploration. *ACL* (2018), 87.

[35] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW*.

[36] Kazunari Sugiyama and Min-Yen Kan. 2010. Scholarly paper recommendation via user's recent research interests. In *JCDL*. 29–38.

[37] Jie Tang, Alvis CM Fong, Bo Wang, and Jing Zhang. 2011. A unified probabilistic framework for name disambiguation in digital library. *TKDE* (2011).

[38] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *SIGKDD*.

[39] C. Tillmann and H. Ney. 2003. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics* 29 (2003), 97–133.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[41] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.

[42] Howard D White, H Cooper, LV Hedges, et al. 2009. Scientific communication and literature retrieval. *The handbook of research synthesis and meta-analysis* 2 (2009), 51–71.

[43] Kan Wu, Jie Tang, Zhou Shao, Xinyi Xu, Bo Gao, and Shu Zhao. 2018. CareerMap: visualizing career trajectory. *Science China Information Sciences* (2018).

[44] Kan Wu, Jie Tang, and Chenhui Zhang. 2018. Where Have You Been? Inferring Career Trajectory from Academic Social Network.. In *IJCAI*.

[45] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).

[46] Zaihan Yang and Brian D Davison. 2012. Venue recommendation: Submitting your paper with style. In *ICMLA*, Vol. 1. IEEE, 681–686.

[47] Jihang Ye, Zhe Zhu, and Hong Cheng. 2013. What's your next move: User activity prediction in location-based social networks. In *SDM*. SIAM, 171–179.

[48] Da Yin, Weng Lam Tam, Ming Ding, and Jie Tang. 2021. MRT: Tracing the Evolution of Scientific Publications. *TKDE* (2021).

[49] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale heterogeneous entity graphs. In *SIGKDD*.

[50] Fanjin Zhang, Jie Tang, Xueyi Liu, Zhenyu Hou, Yuxiao Dong, Jing Zhang, Xiao Liu, Ruobing Xie, Kai Zhuang, Xu Zhang, et al. 2021. Understanding WeChat User Preferences and "Wow" Diffusion. *TKDE* (2021).

[51] Jian-Guo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Xiuming Pan, Yu Gong, and Philip S Yu. 2019. Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce. *NAACL HLT 2019* (2019), 64–72.

[52] Minjia Zhang and Yuxiong He. 2020. Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping. *arXiv preprint arXiv:2010.13369* (2020).

[53] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In *WWW (WWW '21)*. ACM, New York, NY, USA, 3246–3257.

[54] Yutao Zhang, Fanjin Zhang, Peiran Yao, and Jie Tang. 2018. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.. In *SIGKDD*.

[55] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129* (2019).

A EXPERIMENT SUPPLEMENTARY

A.1 Zero-Shot Inference

Use of Prompt Word As shown in Table 5, the use of proposed prompt words in the FOS inference task, turns out to be fairly useful for SciBERT to decode paper fields (FOS). We conjecture it is because the extra appended prompt words can help alter the focus of the pre-training model while making predictions on masked tokens. However, the improvement for SciBERT is marginal on affiliation inference. When decoding venue, it even hurts the performance. This is probably due to the improper choice of prompt words.

For OAG-BERT, this technique has limited help as our expectation. Instead of using continuous positions as SciBERT, OAG-BERT encodes inter-entity positions to distinguish different entities and paper texts. Thus the additional appended prompt word is treated as part of the paper title and is not adjacent to the masked entities for OAG-BERT.

Use of Abstract The use of abstracts can greatly improve the model inference performance in both SciBERT and OAG-BERT. Both models frequently accept long text inputs in the pre-training process, which makes them naturally favor abstracts. Besides, abstracts contain rich text information which can help the pre-training model capture the main idea of the whole paper.

Task Comparisons The affiliation generation task appears to be much harder than the other two tasks. This is probably due to the weak semantic information contained in affiliation names. The words in field-of-studies can be seen as sharing the same language with paper contents and most venue names also contain informative concept words such as "Machine Learning" or "High Energy". This is not always true for affiliation names. For universities like "Harvard University" or "University of Oxford", their researchers could focus on multiple unrelated domains which are hard for language models to capture. For companies and research institutes, some may focus on a single domain but it is not necessary to have such descriptions in their names, which also confuses the pre-training language model.

Discussion for Entity Probability In Equation 2, we use the sum of log probabilities of all tokens to calculate the entity log probability. This method seems unfair for entities with longer lengths as the log probability for each token is always negative. However, for MLM-based models, the encoding process not only encodes "[MASK]" tokens but also captures the length of the masked entity and each token's position. Therefore, if the pre-training corpus has fewer long entities than short entities, in the decoding process, the decoded tokens in a long entity will generally receive higher probability, compared to the ones in a short entity.

Even so, the sum of log probabilities is still not necessary to be the best choice depending on the entity distribution in the pre-training corpus. We conduct a simple experiment to test different average methods. We reform the calculation of entity log probability in Equation 2 as $\frac{1}{L^\alpha}\sum_{1\le k\le l}\log P(w_{i_k}|C, w_{i_1}, w_{i_2}, ..., w_{i_{k-1}})$, where L denotes the length of the target entity. When $\alpha = 0$, this equation degrades to the summation version used in previous tasks. When $\alpha = 1$, this equation degrades to the average version.

We compare different averaging methods by using various α and test their performance on the zero-shot inference tasks. We select the input features with the best performance according to Table 5. For SciBERT, we use both abstract and prompt word for FOS and affiliation inference. We do not use the prompt word for venue inference. For OAG-BERT, we only use abstract as the prompt word does not work well. The results in Table 8 show that for the most time, using the summation strategy outperforms the average strategy significantly. The simple average ($\alpha = 1$) appears to be the worst choice. However, for some situations, a moderate average ($\alpha = 0.5$) might be beneficial.

Table 8: The results for using different average methods while calculating entity log probabilities. Hit@1 and MRR are reported.

Method	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SciBERT			
FOS	**35.33%, 0.52**	32.07%, 0.51	14.85%, 0.36
Venue	18.00%, 0.32	**19.30%, 0.33**	7.07%, 0.23
Affiliation	**12.40%, 0.25**	10.83%, 0.23	9.23%, 0.21
OAGBERT			
FOS	**49.59%, 0.67**	48.08%, 0.66	45.36%, 0.63
Venue	**39.00%, 0.57**	38.20%, 0.57	36.13%, 0.55
Affiliation	**21.67%, 0.38**	19.90%, 0.36	16.47%, 0.31

Discussion for Decoding Order In our designed decoding process, we do not strictly follow the left-to-right order as used in classical decoder models. The main reason is that for encoder-based BERT model, the decoding for each masked token relies on all bidirectional context information, rather than only prior words. We compare the performance of using left-to-right decoding and out-of-order decoding in Table 9.

The results show that for FOS, there is no significant difference between two decoding orders, since the candidate FOS only has one or two tokens inside. As for venue and affiliation, it turns out that the out-of-order decoding generally performs much better than left-to-right decoding, except when OAG-BERT uses abstract where differences are relatively small as well. We also present the results for models using left-to-right decoding and prompt words in Table 9, which indicates that the left-to-right decoding will sometimes undermine the effectiveness of prompt words significantly, especially for OAG-BERT.

Table 9: The results for using left-to-right decoding and out-of-order decoding order. Hit@1 and MRR are reported. Results with difference larger than 1% Hit@1 were bolded.

Method	FOS		Venue		Affiliation	
	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR
SciBERT						
Left-to-Right	20.05%	0.37	8.40%	0.20	6.90%	0.18
Out-of-Order	19.93%	0.37	**9.87%**	**0.22**	6.93%	0.19
SciBERT *+prompt*						
Left-to-Right	29.65%	0.47	9.57%	0.21	8.03%	0.20
Out-of-Order	29.59%	0.47	10.03%	0.21	8.00%	0.20
SciBERT *+abstract*						
Left-to-Right	25.67%	0.43	11.43%	0.24	7.63%	0.19
Out-of-Order	25.66%	0.43	**18.00%**	**0.32**	**10.33%**	**0.22**
SciBERT *+both*						
Left-to-Right	35.21%	0.52	**11.17%**	**0.24**	11.47%	0.23
Out-of-Order	35.33%	0.52	9.83%	0.22	12.40%	0.25
OAG-BERT						
Left-to-Right	34.94%	0.53	11.33%	0.24	5.47%	0.17
Out-of-Order	34.36%	0.51	**21.00%**	**0.37**	**11.03%**	**0.24**
OAG-BERT *+prompt*						
Left-to-Right	37.84%	0.56	12.53%	0.26	5.50%	0.17
Out-of-Order	37.33%	0.55	**22.67%**	**0.39**,	**11.77%**	**0.25**
OAG-BERT *+abstract*						
Left-to-Right	49.75%	0.67	**40.50%**	**0.59**	21.93%	0.38
Out-of-Order	49.59%	0.67	39.00%	0.57	21.67%	0.38
OAG-BERT *+both*						
Left-to-Right	49.83%	0.67	22.17%	0.38	6.80%	0.19
Out-of-Order	49.51%	0.67	**38.47%**	**0.57**	**21.53%**	**0.38**

Table 10: A full list of used candidates in zero-shot inference tasks and supervised classification tasks.

FOS: Art, Biology, Business, Chemistry, Computer science, Economics, Engineering, Environmental science, Geography, Geology, History, Materials science, Mathematics, Medicine, Philosophy, Physics, Political science, Psychology, Sociology
Venue: Arxiv: algebraic geometry, Arxiv: analysis of pdes, Arxiv: astrophysics, Arxiv: classical analysis and odes, Arxiv: combinatorics, Arxiv: computer vision and pattern recognition, Arxiv: differential geometry, Arxiv: dynamical systems, Arxiv: functional analysis, Arxiv: general physics, Arxiv: general relativity and quantum cosmology, Arxiv: geometric topology, Arxiv: group theory, Arxiv: high energy physics - experiment, Arxiv: high energy physics - phenomenology, Arxiv: high energy physics - theory, Arxiv: learning, Arxiv: materials science, Arxiv: mathematical physics, Arxiv: mesoscale and nanoscale physics, Arxiv: nuclear theory, Arxiv: number theory, Arxiv: numerical analysis, Arxiv: optimization and control, Arxiv: probability, Arxiv: quantum physics, Arxiv: representation theory, Arxiv: rings and algebras, Arxiv: statistical mechanics, Arxiv: strongly correlated electrons
Affiliation: Al azhar university, Bell labs, Carnegie mellon university, Centers for disease control and prevention, Chinese academy of sciences, Electric power research institute, Fudan university, Gunadarma university, Harvard university, Ibm, Intel, Islamic azad university, Katholieke universiteit leuven, Ludwig maximilian university of munich, Max planck society, Mayo clinic, Moscow state university, National scientific and technical research council, Peking university, Renmin university of china, Russian academy of sciences, Siemens, Stanford university, Sun yat sen university, Tohoku university, Tsinghua university, University of california berkeley, University of cambridge, University of oxford, University of paris

Table 11: The sizes for datasets used in supervised classification tasks.

Task	Categories	Train	Validation	Test
FOS	19	152000	19000	19000
Venue	30	24000	3000	3000
Affiliation	30	24000	3000	3000

Table 12: Details for the CS heterogeneous graph used in the link prediction.

	Papers	Authors	FOS
Nodes	544244	510189	45717
1116163	Venues	Affiliations	
	6934	9079	
	#Paper-Author	#Paper-FOS	#Paper-Venue
#Edges	1862305	2406363	551960
6389083	#Author-Affiliation	#Paper-Paper	#FOS-FOS
	519268	992763	56424