

OAG-BERT: Pre-train Heterogeneous Entity-augmented Academic Language Models

Xiao Liu^{†*}, Da Yin^{†*}, Xingjian Zhang[†], Kai Su[†], Kan Wu[†], Hongxia Yang[‡], Jie Tang^{†**}

[†] Department of Computer Science and Technology, Tsinghua University, China

[‡] DAMO Academy, Alibaba Group, China

{liuxiao17,yd18}@mails.tsinghua.edu.cn,yang.yhx@alibaba-inc.com,jietang@tsinghua.edu.cn

ABSTRACT

To enrich language models with domain knowledge is crucial but difficult. Based on the world’s largest public academic graph Open Academic Graph (OAG), we pre-train an academic language model, namely OAG-BERT, which integrates massive heterogeneous entities including paper, author, concept, venue, and affiliation. To better endow OAG-BERT with the ability to capture entity information, we develop novel pre-training strategies including heterogeneous entity type embedding, entity-aware 2D positional encoding, and span-aware entity masking. For zero-shot inference, we design a special decoding strategy to allow OAG-BERT to generate entity names from scratch. We evaluate the OAG-BERT on various downstream academic tasks, including NLP benchmarks, zero-shot entity inference, heterogeneous graph link prediction and author name disambiguation. Results demonstrate the effectiveness of the proposed pre-training approach to both comprehending academic texts and modeling knowledge from heterogeneous entities. OAG-BERT has been deployed to multiple real-world applications, such as reviewer recommendations and paper tagging in the AMiner system. OAG-BERT¹ is also available to the public through the CogDL package.

KEYWORDS

Pre-training, Language Modeling, Knowledge Representation, Heterogeneous Graph

1 INTRODUCTION

Pre-trained language models such as BERT [8], GPT [36] and XL-Net [44] substantially promote the development of natural language processing. Besides pre-training for general purposes, more and more language models are targeting at specific domains, such as BioBERT [25] for biomedical field and SciBERT [2] for academic field, which establish new state-of-the-art on many domain-related benchmarks such as named entity recognition [9, 31], topic classification [4, 18] and so on.

However, most of these models are only pre-trained over domain corpus, but ignore to integrate domain entity knowledge, which is crucial for many entity-related downstream tasks. In the author name disambiguation task, the affiliation of a paper could contribute by indicating the field-of-study of an author. For example, authors from Max Planck Institute may focus more on science and engineering rather than humanity. We may also produce fine-grained

¹<https://github.com/thudm/oag-bert>

*These authors contributed equally to this work.

**Jie Tang is the corresponding author.

OAG-BERT: Pre-train Heterogeneous Entity-augmented Academic Language Models

Xiao Liu

中国清华大学计算机科学与技术系

Tang^{†**}

[‡]

{liuxiao17,yd18}@mails.tsinghua.edu.cn,yang.yhx@alibaba-inc.com,jietang@tsinghua.edu.cn

ABSTRACT

为了丰富具有域知识的语言模型至关重要但困难。基于世界上最大的公众学术图公开学术图（OAG），我们预先培训了学术语言模型，即OAG-BERT，它整合了包括纸张，作者，概念，场地和隶属关系的大规模异构侵权。为了更好地获得捕获实体信息的能力，我们开发了新的预训练策略，包括异构实体类型嵌入，实体感知2D位置编码和跨度感知实体屏蔽。对于零拍摄推断，我们设计了一种特殊的解码策略，以允许OAG-BERT从头开始生成姓名。我们在各种下游学术任务中评估OAG-BERT，包括NLP基准，零射实体推断，异构图链路预测和作者名称消歧。结果表明，拟议的预训练方法对理解学术文本和从异质实体建模知识的有效性。OAG-BERT已部署到多个现实世界的应用程序，例如散对系统中的审阅者建议和纸质标记。OAG-BERT1还可以通过COGDL包可供公众使用。

KEYWORDS

预培训，语言建模，知识表示，异构图

1 INTRODUCTION

预先接受的语言模型如BERT [8]，GPT [36]和XL-Net [44]大大提升了自然语言处理的发展。除了进行普通目的的预先培训外，越来越多的语言模型是针对特定领域的针对性，例如生物医学领域的Biobert [25]，以及学术领域的Scibert [2]，在许多域中建立新的最先进 – 命名实体识别[9,31]，主题分类[4,18]等重写基准。

但是，大多数这些模型仅在域中预先培训语料库，但忽略整合域实体知识，这对于许多与许多与实体相关的下游任务至关重要。在作者名称消歧任务中，纸张的隶属度可以通过表明作者的研究领域来贡献。例如，来自Max Planck Institute的作者可以更多地关注科学和Engi-Leenering而不是人性。我们也可能产生细粒度

¹<https://github.com/thudm/oag-bert>

*这些作者同样为这项工作贡献。**杰唐是相应的作者。

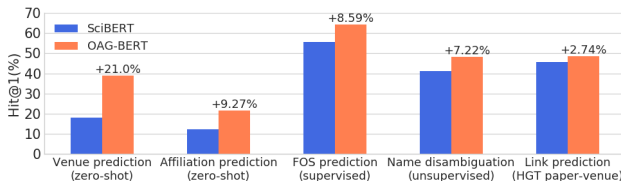


图1：OAG-BERT以2.75%–21.0%（绝对增益）的一系列实体相关任务的eag-bert优于SciBERT。

通过利用来自标题，抽象，附属机构甚至其作者名称的信息，从数据库中进行研究。这将大大提高研究人员的有效性，以确定其相关的兴趣工作。

作为对语料库的补充，许多域知识图表

可提供知识。例如，OAG

[46]是载有超过7亿实体的最大公众的异构学术实体图，包括纸张，作者，学习，场地和隶属;和20亿个关系。如果我们可以将如此巨大的知识纳入域语言模型，则会受益许多下游应用。但是，尽管有丰富的知识来源，如何将知识注入语言模型已成为关键问题。虽然许多作品一直专注于从大规模知识图中注入同质实体和关系知识[35,39,50]，但没有一个人认为异构场景的问题通常在实践中持有真实。

要弥补差距，我们向共同模本文本提出了OAG-BERT通过在OAG上预先训练，通过预培训而异的实体知识。我们收集约500万条全文和1.1亿摘要作为Corpora，7000万纸自我网络，包括其作者，学习领域，场地和附属机构。为了处理异核，我们分别为每种类型的实体设计异构实体类型嵌入式。要实现具有各种长度的实体名称的屏蔽语言，我们设计了一种新颖的Spann感知实体屏蔽策略，可以根据实体长度选择屏蔽令牌的连续跨度。为了更好地使用实体跨度和序列顺序“注意”OAG-BERT，我们提出了实体感知的2D位置编码，以考虑实体间序列顺序和实体内部令牌顺序。

我们首先在各种类型的下游评估OAG-BERT应用程序，包括传统学术NLP数据集[2,5]，新型零射实体推理[34]，异构图学习（链路预测）[10,16]和作者名称消歧[3,49]。对于零拍摄推断，我们开发了一种特殊的解码策略

for OAG-BERT, allowing it to generate fluent sequences like GPTs. Not only do the experiment results demonstrate OAG-BERT’s competitive performance to previous pre-trained models on ordinary language benchmarks, but also support its outstanding grasp of entity knowledge by outperforming over tasks that heavily depend on entity knowledge.

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

- We propose to study the problem of enriching pre-trained language models with heterogeneous entity knowledge. To solve the problem, we design heterogeneous entity type embedding, span-aware entity masking and entity-aware 2D positional encoding. We also develop a special decoding strategy for BERT-style models to generate high-quality entities from scratch.
- We present the OAG-BERT, an entity knowledge augmented academic language model that is pre-trained over 5 million paper full-text, 110 million paper abstracts and billions of academic entities and relations from the OAG. It has the similar number of parameters with other BERT-based models such as SciBERT.
- We conduct relatively extensive experiments to demonstrate OAG-BERT’s capability of traditional language tasks, zero-shot inference, heterogeneous graph learning and author name disambiguation. OAG-BERT has been deployed as the infrastructure of AMiner system² for OAG downstream applications. It is open to the public access through the CogDL [45] package.
- We apply the pre-trained OAG-BERT model to several real-world applications, such as the reviewer recommendation. It is also employed as a fundamental component in the AMiner system, which is further used to improve the performance on tasks like automatic paper tagging or author name disambiguation.
- We release the pre-trained OAG-BERT model in CogDL package for open access and free use.

2 RELATED WORKS

Our proposed OAG-BERT model is based on BERT [8], a self-supervised [28] bidirectional language model. It employs multi-layer transformers as its encoder and uses masked token prediction as its objective, which allows using massive unlabeled text data as training corpus. The model architecture and training scheme have been shown to be effective on various natural language tasks, such as question answering or natural language inference.

BERT has many variants. Some focus on the robustness of the pre-training process, like RoBERTa [29]. Some others try to incorporate more knowledge into the natural language pre-training. SpanBERT [17] develops span-level masking which benefits span selection tasks. ERNIE [50] introduces explicit knowledge graph inputs to the BERT encoder and achieves significant improvements over knowledge-driven tasks.

As for the academic domain, previous works such as BioBERT [25] or SciBERT [2] leverage the pre-training process on scientific domain corpus and achieve state-of-the-art performance on several academic NLP tasks. The S2ORC-BERT [30], applies the same method with SciBERT on a larger scientific corpus and slightly improves the performance on downstream tasks. Later works [14] 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 [1]. Other large academic corpora including AMiner [40], OAG [40, 46], and Microsoft Academic Graph (MAG) [19] also integrate massive publications with rich graph information as well, such as authors and research fields.

On academic graphs, there are some tasks that involve not only text information from papers but also structural knowledge lying behind graph links. For example, to disambiguate authors with the same names [3, 49], the model needs to learn node representations in the heterogeneous graph. To better recommend papers for online academic search [11, 12], graph information including related academic concepts and published venues could provide great benefits. To infer experts’ trajectory across the world [43], 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.

Targeting at graph-based problems, many graph representation learning methods were proposed in the last decade. Works like node2vec [13] and ProNE [47] focus on purely homogeneous graph structures and metapath2vec [10] later extends the idea to heterogeneous graphs. Neural-based methods like GCN [23] successfully introduce neural networks to solve the graph learning problem. Recent works including Heterogeneous Graph Transformer [16] and GPT-GNN [15] similarly borrow the idea from the natural language community, applying transformer blocks and pre-training scheme on graph tasks.

3 METHODS

The proposed OAG-BERT is a bidirectional transformer-based pre-training model. It can encode scientific texts and entity knowledge into high dimensional embeddings, which can be used for downstream tasks such as predicting the published venue for papers. We build the OAG-BERT model on top of the conventional BERT [8] model with 12 transformer [42] encoder layers.

While the original BERT model only focuses on natural language, our proposed OAG-BERT also incorporates heterogeneous entity knowledge. In other words, in addition to learning from pure scientific texts such as paper title or abstract, the OAG-BERT model can comprehend other types of information, such as the published venues or the affiliations of paper authors. To achieve that, we made several modifications to the model architecture and the pre-training process. We will introduce them in the following sections. An overview of the proposed OAG-BERT model is depicted in Figure 2.

3.1 Model Architecture

The key challenge for OAG-BERT lies in how to integrate knowledge into language models. Previous approaches [27, 50] mainly focus on injecting homogeneous entities and relations from knowledge graphs like Wikidata, and very few of them look into situations where there are heterogeneous entities.

To augment OAG-BERT with various types of entity knowledge, we place title, abstract along other entities from the same paper in a single sequence as one training instance (see Figure 2).

对于OAG-BERT，允许它生成像GPT的流畅序列。实验结果不仅可以证明OAG-BERT对普通语言基准测试的先前预先训练的型号，而且通过大量依赖实体知识的任务，还支持其出色的对实体知识的掌握。

总而言之，我们在本文中提出以下贡献：

我们建议研究丰富训练的LAN的问题 –

具有异构实体知识的贡献模型。为了解决问题，我们设计异构实体类型嵌入，Spanive Intenty屏蔽和实体感知2D位置编码。我们还开发了一种特殊的解码策略，可为BERT风格的MOD-ELS生成从头开始产生高质量实体。

我们介绍了OAG-BERT，一个实体知识增强

学术语言模型，预先培训超过500万纸全文，1.1亿纸摘要和数十亿的学术实体和奥格关系。它具有与其他基于BERT的型号相似的参数，例如SCIBERT。

我们进行相对广泛的实验来证明

OAG-BERT的传统语言任务的能力，零射垒推断，异构图形学习和作者名称不违反 – 大。OAG-BERT已被部署为AMINS System2的基础设施，用于OAG下游应用程序。通过Cogdl [45]包装，公众访问是开放的。

我们将预先训练的OAG-BERT模型应用于几个真实世界申请，例如审阅者建议。它还被用作aminer系统中的基本组件，该组件还用于改善自动纸标记或作者名称消歧的任务的性能。

我们在Cogdl包中释放了预先训练的OAG-BERT模型开放访问和免费使用。

2 RELATED WORKS

我们提出的OAG-BERT模型基于BERT [8]，自我监督的[28]双向语言模型。它使用多层变压器作为其编码器，并使用屏蔽令牌预测作为其目标，允许使用大规模未标记的文本数据作为培训语料库。模型架构和培训方案已被证明对各种自然语言任务有效，例如问题应答或自然语言推断。

伯特有很多变体。一些专注于鲁棒性预训练过程，如罗伯塔[29]。有些其他人试图将更多知识纳入自然语言预训练中。Spanbert [17]开发跨度级屏蔽，它有益于跨度选择任务。ernie [50]向BERT编码器引入显式知识图输入，并通过知识驱动的任务实现了显著的改进。

至于学术领域，以前的作品如

Biobert [25]或Scibert

[2]利用科学域语料库的预培训过程，实现了几个学术NLP任务的最先进的性能。S2ORC-BERT [30]将与频闪相同的方法应用于更大的科学语料库，略微提高下游任务的性能。后来作品[14]进一步表明，对特定的组件的持续培训也有利于下游任务。这些学术界

培训模型依赖大型科学集团。Scibert使用语义学者语料库[1]。其他大型学术集团包括aminer [40]，oag [40,46]和Microsoft学术图（Mag）[19]还将大规模出版物与丰富的图形信息相结合，如作者和研究领域。

在学术图上，有一些不仅涉及的任务

来自论文的文本信息，但也是躺在图表链接后面的结构知识。例如，要歧视具有相同名称[3,49]的作者，模型需要学习异构图中的节点表示。为了更好地推荐在线学术搜索文件[11,12]，图表信息包括相关的ACA-Demic概念和公布的场地可以提供巨大的好处。在全球推断专家的轨迹[43]，将AU与他们的联系与语义层面相关联。捕获纸张标题或摘要的功能远非足够的这些类型的挑战。

针对基于图形的问题，许多图表表示

在过去十年中提出了学习方法。像Node2vec [13]这样的工作，易于[47]专注于纯粹的均匀图形结构和Metapath2Vec [10]后来将想法延伸到异质图。基于神经的方法，如GCN [23]成功地引入了神经网络以解决图表学习问题。包括异构图形变压器的重复工作[16]和GPT-GNN [15]类似地借用自然语言社区的想法，在图表任务上应用变压器块和预培训方案。

3 METHODS

所提出的OAG-BERT是一种基于双向变压器的预训练模型。它可以将科学文本和实体知识编码为高维嵌入式，这可以用于降低流动任务，例如预测纸张的已发布的场地。我们在带有12个变压器[42]编码器层的传统BERT [8]型号顶部构建OAG-BERT模型。

虽然原始BERT型号仅重点关注自然LAN –

我们提出的OAG-BERT还包含异质实体知识。换句话说，除了从纸张标题或抽象等纯科学文本学习之外，OAG-BERT模型可以理解其他类型的信息，例如有趣的场所或纸作者的附属物。为此，我们对模型架构和预培训过程进行了多次修改。我们将在以下分区中介绍它们。图2中描绘了所提出的OAG-BERT模型的概述。

3.1 Model Architecture

OAG-BERT的关键挑战在于如何将知识集成到语言模型中。以前的方法[27,50]主要专注于注入均匀的实体和与Wikidata这样的知识图中的关系，并且很少有人看出存在异构实体的情况。

使用各种类型的实体知识来增强OAG-BERT

我们在单个序列中划分标题，摘要沿其他纸张的其他实体作为一个训练实例（参见图2）。

²<https://www.aminer.cn>

²<https://www.aminer.cn>

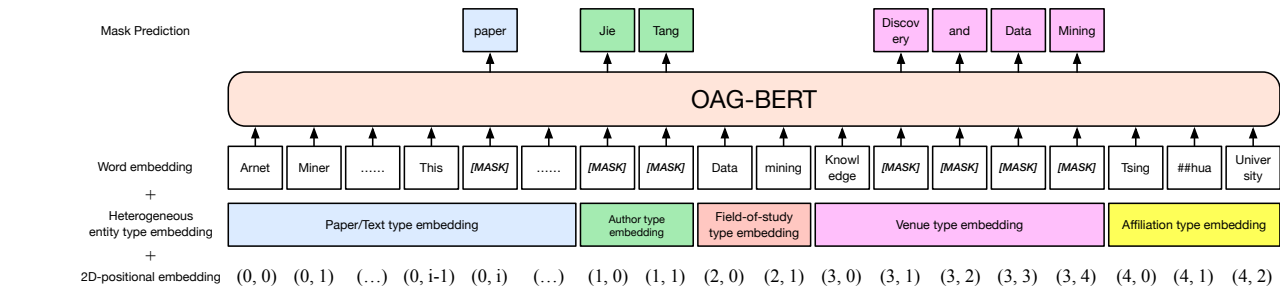


Figure 2: Heterogeneous entity augmentation in OAG-BERT. 1) For different entity types, we design heterogeneous entity type embedding. 2) For comparatively long entities (such as the “Knowledge Discovery and Data Mining”), we leverage the span-aware entity masking strategy, which selects a continuous span in the entity. 3) For positional embeddings across different entities, we design an entity-aware 2D-positional embedding strategy, whose first dimension is designed for indicating inter-entity sequence order, and the second dimension is designed for indicating intra-entity token sequence order.

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 rest four types of entities. Following the notation in OAG, we use FOS (field-of-study) to denote research fields. Thanks to OAG, for venues, authors, affiliations, and FOS, their names have been cleaned up, deduplicated, and unified, which enables OAG-BERT to learn a consistent representation for each entity.

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. The original BERT employs the next sentence prediction loss (NSP) to learn the relationship between sequences, which requires the use of token type embeddings to distinguish two sequences from each other. Tokens from two sequences are added by different token type embeddings. However, NSP loss is believed to harm rather than improve BERT’s performance, as found in later works [17, 29]. Therefore, in this work, we abandon the NSP loss and discard the old token type embeddings.

On the other hand, 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 are 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. Although the transformer [42] architecture has achieved great success in sequence-based tasks, it is also known that the transformer itself is

permutation-invariant (i.e. is not aware of the sequence order). The critical technique of applying transformer to natural language is to add a *positional embedding* to indicate the sequence order, including the absolute positional embedding used in vanilla Transformer [42] and BERT [8], and the relative positional embedding developed in Transformer-XL [6] and XLNet [44].

However, when we want OAG-BERT to capture entity knowledge, neither of them is applicable. This is because the conventional positional embedding can not distinguish words from entities that are adjacent to each other and of the same type. For instance, if there are two affiliations “Tsinghua University” and “Unviersity 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 to 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 (see Figure 2). The first dimension is for inter-entity order, indicating the token is in which entity; 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 title and abstract, we adopt the same random masking strategy as in BERT. However, for entities such as author names, field of study, venues and affiliations, to encourage OAG-BERT to memorize them, we develop a span-aware entity masking strategy which combines the advantages of both ERNIE [50] and SpanBERT [17].

The intuition of using this strategy is that, some of the entities are too long and thus too difficult for the OAG-BERT to learn. The span-aware entity masking strategy not only alleviates the problem, but also still preserves the sequential relationship of an entity’s tokens: for entity that has less than 4 tokens, we will mask the whole entity; and for others, we sample masked lengths from a

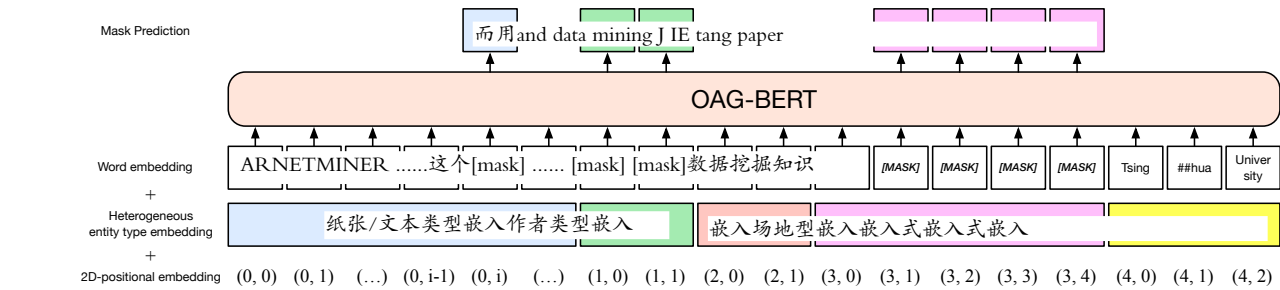


图2：OAG-BERT中的异构实体增强。1）对于不同的实体类型，我们设计异构实体类型嵌入。2）对于相对较长的实体（例如“知识发现和数据挖掘”），我们利用了Spanive Invent的实体屏蔽策略，该屏蔽策略在实体中选择连续跨度。3）对于跨不同实体的位置嵌入，我们设计一个实体感知的2D位置嵌入策略，其第一维度被设计用于指示实体界序列顺序，并且第二维设计用于指示实体内令牌序列顺序。

总共有五种类型的实体。我们对待文本功能纸张的（标题和摘要）作为一个特殊文本实体。有趣的场地，作者，隶属关系以及研究领域是其余四种类型的实体。在OAG中的符号之后，我们使用FOS（学习场）来表示研究领域。由于OAG，对于场地，作者，附属机构和FOS，他们的名字已经清理了，重复数据删除和统一，这使OAG-BERT能够为每个实体学习一致的代表。

从一篇纸上的所有实体都作为输入衔接样本。为了帮助OAG-BERT模型区分它们，我们使用另外三种技术：异构实体类型嵌入，实体感知2D位置编码和跨度感知实体屏蔽。

异构实体类型嵌入。原来的伯特采用下一个句子预测丢失（NSP）来学习序列之间的相关性，这需要使用令牌类型嵌入来区分彼此的两个序列。来自两个序列的令牌由不同的令牌类型嵌入式添加。然而，据信NSP损失涉及危害而不是提高BERT的性能，如后面的作品所发现的[17,29]。因此，在这项工作中，我们放弃了NSP丢失并丢弃了旧令牌类型的嵌入式。

另一方面，为了区分不同类型的实体，我们建议利用在预训练过程中嵌入实体类型以指示实体类型，其使用类似于BERT中使用的令牌类型嵌入。

例如，鉴于纸张的标题和摘要“Arnetminer：学术社交网络的提取和挖掘”，我们检索其作者，研究，场所和附属实体，并将它们连接到少于512令牌的序列中。对于纯文本（例如标题和摘要），我们将它们标记为原始实体类型索引（例如，0）以获取其实体类型嵌入。对于作者实体（如jie jang），我们使用作者类型索引标记它们（例如，1）。所以对于其他实体来说。是什么，因为实体在序列中是订单不变的，我们在样本序列中抽他的顺序，以避免我们的模型来学习这些实体的任何位置偏差。

实体感知2D位置编码。虽然跨越前[42]架构在基于序列的任务中取得了巨大的成功，也众所周知，变压器本身是

置换不变（即不知道序列顺序）。将变压器应用于自然语言的临界技术是添加一个位置嵌入，以指示序列顺序，包括在香草转换器[42]和BERT [8]中使用的绝对位置嵌入，以及在变压器中产生的相对位置嵌入 - XL [6]和XLNET [44]。

但是，当我们希望oag-bert捕获实体知识边缘，它们都不适用。这是因为传统的位置嵌入不能区分与彼此相邻的实体和相同类型的单词。例如，如果有两个“清华大学”和“加利福尼亚州的不佛罗里州”在一系列中互相放置，那么传递者将认为有一个名为“清华大学加利福尼亚州大学”的隶属关系。

总而言之，我们的请求可以总结到两点：1）位置嵌入应该暗示实体间序列顺序（用于区分不同实体的）和2），所谓的嵌入应该指示实体内令牌序列顺序（其用作传统的位置嵌入）。

鉴于此，我们设计实体感知的2D位置床上用品解决实体间和实体内部问题（见图2）。第一个维度用于实体间顺序，指令是在哪个实体中;第二维为实体内的顺序，指示令牌的序列。对于给定位置，通过将两个位置嵌入在一起来计算最终位置嵌入。

跨度感知实体屏蔽。在执行屏蔽时，对于纯文本内容（如纸张标题和摘要），我们采用与伯特相同的随机掩蔽策略。然而，对于诸如作者名称，学习领域，场所和附属机构等实体，以鼓励OAG-BERT来记住它们，我们开发了一个跨度感知的实体掩蔽策略，它结合了ernie [50]和spanbert [17]的优势。

使用这种策略的直觉是，一些实体太长时间了，因此oag-bert来说太难了解。跨越的实体屏蔽策略不仅可以减轻问题，而且还保留了实体令牌的顺序关系：对于具有少于4个令牌的实体，我们将掩盖整个实体;对于其他人来说，我们从a中屏蔽长度

geometric distribution $\text{Geo}(p)$ which satisfies:

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

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 [37], where layer normalization is placed inside the residual connection instead of after the add-operation in Transformer blocks. Previous work [48] demonstrates that training with Pre-LN BERT avoids vanishing gradients when using aggressive learning rates. Therefore, it is shown to be more stable than the traditional Post-LN version for optimization.

3.2 Pre-training 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 title, abstract, venue, authors, affiliations, and field-of-studies (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, which mainly consists of arXiv PDF dumps; the other comes from the PubMed XML dump. We clean up and sentencize the corpus with SciSpacy [32]. 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 original BERT implementation. This ends up with 44,000 unique tokens in our vocabulary.

For better handling the entity knowledge of authors in the OAG, in the data preprocessing we transform the author name list as a sentence for each paper and place it between the title and abstract. 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. After the loss has converged, we 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 picked all authors with at least 3 papers published. Then we filtered 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.

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 EXPERIMENTS

In this section, we will introduce several experiments to demonstrate the effectiveness of our proposed OAG-BERT. First, to exhibit how OAG-BERT works on multi-type information, we design intuitive zero-shot inference tasks. Then, we make extensions to supervised classification tasks. We further apply the pre-trained embeddings to name disambiguation and link prediction tasks, which present the superior capability of OAG-BERT in leveraging various types of entities. Finally, on the NLP tasks used by SciBERT, we additionally verify that the proposed OAG-BERT model can also achieve competitive results with text-only information provided. An overview of the model performance is shown in Table 1.

4.1 Zero-shot Inference

Although not using 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 the pre-training process. We develop a simple extension to the Masked Language Model (MLM) to achieve that.

In MLM, the token prediction task in the pre-training process can be seen as maximizing the probability of masked input tokens. It treats the predictions for each token as independent processes. The target can be denoted as maximizing $\sum_{w \in \textit{masked}} \log P(w|C)$, where *masked* is the collection of masked tokens and *C* denotes contexts, which represents the inputs of MLM, including both input tokens and position information.

In the entity decoding process, we cannot ignore the dependencies between tokens in each entity, which requires us 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

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

However, the number of possible decoding orders is *l*!, which makes it extremely expensive to calculate while dealing with long entities. Thus, we adopt two strategies to solve this problem. First, while calculating the probability for one given entity, we use greedy selection to decide the decoding order. In other words, for each round of decoding, we choose the token with maximal probability to decode. An example is depicted in Figure 3. Second, when decoding an entity from scratch, we use beam search [41] to search the token combinations with the highest probability.

Another challenge for decoding using the MLM model is to choose the appropriate entity length. Instead of using fixed length

geometric distribution $\text{Geo}(p)$ which satisfies:

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

如果采样的长度小于实体长度，我们只会掩盖实体。对于文本内容和实体内容，我们将分别掩盖15%的令牌。

Pre-LN Bert.除了以前的原始BERT架构的更改外，我们进一步采用了DeepSpeed [37]中使用的前LN BERT，其中层归一化放置在剩余连接内，而不是在转换器块中的添加操作之后。以前的工作[48]演示了使用Pre-LN BERT的训练在使用激进学习率时避免了消失的梯度。因此，显示比传统的LN版本更稳定，以进行优化。

3.2 Pre-training Details

OAG-BERT的预训练分为两个阶段。在第一阶段，我们只使用科学文本（纸张标题，摘要和身体）作为模型输入，而不使用上面介绍的实体增强输入。该过程类似于原始BERT模型的预训练。我们将中间体预先训练的模型命名为oag-bert的香草版。在第二阶段，基于Vanilla OAG-BERT，我们继续培训在异构实体上的模型，包括标题，摘要，场地，作者，附属机构和研究领域（FOS）。

第一阶段：预先训练Vanilla OAG-BERT。在培训前的第一阶段，我们从两个来源构建培训语料库：一个来自aminer的PDF存储，主要由Arxiv PDF转储组成;另一个来自PubMed XML转储。我们清理并将语料库置于SCISPacy [32]。这些语料库增加了多个学科的大约500万个独特的纸张全文。在词汇量方面，我们使用Wordpiece构建我们的OAG-BERT词汇，其也用于原始BERT实现。在我们的词汇中，这最终有44,000个独特的令牌。

为了更好地处理OAG中作者的实体知识，在数据预付款中，我们将作者名称列表转换为每篇论文的句子，并将其放在标题和副本之间。因此，与Perfibert这样的以前的模型相比，我们的词汇包含来自作者名称的更多代币。

在伯特的培训程序之后，香草oag-bert首先预先培训，最多128个令牌。损失融合后，我们转向预先训练它的样品，用512令牌。

第二阶段：用实体知识丰富OAG-BERT。在预培训的第二阶段，我们使用OAG语料库中的论文和相关实体。与第一阶段中使用的语料库相比，我们没有为OAG中的所有文件都有完整的文本。因此，我们只使用纸张标题和摘要作为纸质文本信息。从这个语料库中，我们挑选了至少3篇论文发布的作者。然后我们过滤了与这些所选作者无关的所有文件。最后，我们获得了1.2亿篇论文，1000万作者，67万个人，5.6万个场地和26万个附属机构。按照上述输入构建方法，每张纸张及其连接实体都被连接到单个训练实例中。在这个阶段，我们整合了一节中提到的三种策略

3.1以oag-bert为“注意到”实体的能力，而不是将它们视为纯文本。

我们的预培训是用32个NVIDIA Tesla V100 GPU进行的，累计批量大小为32768。我们使用DeepSpeed中的默认BERT预训练配置。我们为第一阶段进行了16K步，第二阶段的另一级预训练和另外4K步。

4 EXPERIMENTS

在本节中，我们将介绍几个实验，以证明我们提出的OAG-BERT的有效性。首先，要以外的是OAG-BERT如何在多型信息上工作，我们设计直观的零点推理任务。然后，我们向监督分类任务进行扩展。我们进一步应用预训练的em-床位以命名消歧和链接预测任务，这提出了OAG-BERT在利用各种类型的实体方面的优异能力。最后，在Scibert使用的NLP任务上，我们还验证了所提出的OAG-BERT模型还可以通过仅提供文本信息实现竞争结果。模型性能的概述如表1所示。

4.1 Zero-shot Inference

虽然不使用像GPT-3这样的单向解码器结构，但我们发现基于双向编码器的OAG-BERT也能够基于在预训练过程中学到的知识来解码实体。我们开发了一个简单的扩展到蒙版语言模型（MLM）以实现这一目标。

在MLM中，在预培训过程中令牌预测任务可以看出，最大化屏蔽输入令牌的概率。它将每个令牌视为独立进程的预测。目标可以表示为最大化

其中屏蔽是屏蔽令牌和C表示上下文的集合，其表示MLM的输入，包括输入令牌和位置信息。

在实体解码过程中，我们不能忽略依赖 - 每个实体的令牌之间的配置，这要求我们共同考虑一个实体中所有令牌的概率，如下面的日志（ $\frac{1}{c} \prod_{i=1}^l P(w_{i_k}|C, w_{i_1}, w_{i_2}, ..., w_{i_{k-1}})$ ），其中 $\frac{1}{c}$ 是实体长度和 $\prod_{i=1}^l$ 是实体中的令牌。由于MLM不是单向模型，一个实体中令牌的参数顺序可以是任意的。假设解码顺序是 $\frac{1}{c} \prod_{i=1}^l P(w_{i_k}|C, w_{i_1}, w_{i_2}, ..., w_{i_{k-1}})$ ，其中 $\frac{1}{c}$ 是1,2, ..., ..., ..., ...。然后可以将预测目标改革为最大化

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

但是，可能的解码订单的数量是 $l!$ ，这使得在处理长实体的同时计算成本非常昂贵。因此，我们采用两种策略来解决这个问题。首先，在计算一个给定实体的概率时，我们使用贪婪选择来决定解码顺序。换句话说，对于每轮解码，我们选择具有最大概率的令牌来解码。图3中描绘了一个例子。第二，当从划痕解码实体时，我们使用波束搜索[41]来搜索具有最高概率的令牌组合。

使用MLM模型进行解码的另一个挑战是选择合适的实体长度。而不是使用固定长度

Table 1: The summary of model performance for all tasks. We report the performance of only using paper titles as inputs in *title-only* and the best performance of using other features such as FOS or venue as inputs in *mixed*.

Method		Zero-shot Inference ¹			Supervised Classification ²			NA ³	Link Prediction ⁴	
		FOS	Venue	Affiliation	FOS	Venue	Affiliation		Paper-Field	Paper-Venue
SciBERT	<i>title-only</i>	29.59%	10.03%	8.00%	55.13%	61.86%	35.44%	0.3690	0.4740	0.4570
	<i>mixed</i>	35.33%	18.00%	12.40%	55.63%	78.05%	56.04%	0.4101	-	-
OAG-BERT	<i>title-only</i>	37.33%	22.67%	11.77%	54.54%	63.03%	35.04%	0.4120	0.4892	0.4844
	<i>mixed</i>	49.59%	39.00%	21.67%	64.22%	78.47%	57.63%	0.4823	-	-

^{1,2} Hit@1 is reported for zero-shot inference and supervised classification.

³ NA is short for Name disambiguation. The macro pairwise f1 score is reported.

⁴ For link prediction tasks, we use pre-trained models to encode all types of nodes. Only title was provided for paper nodes. NDCG is reported.

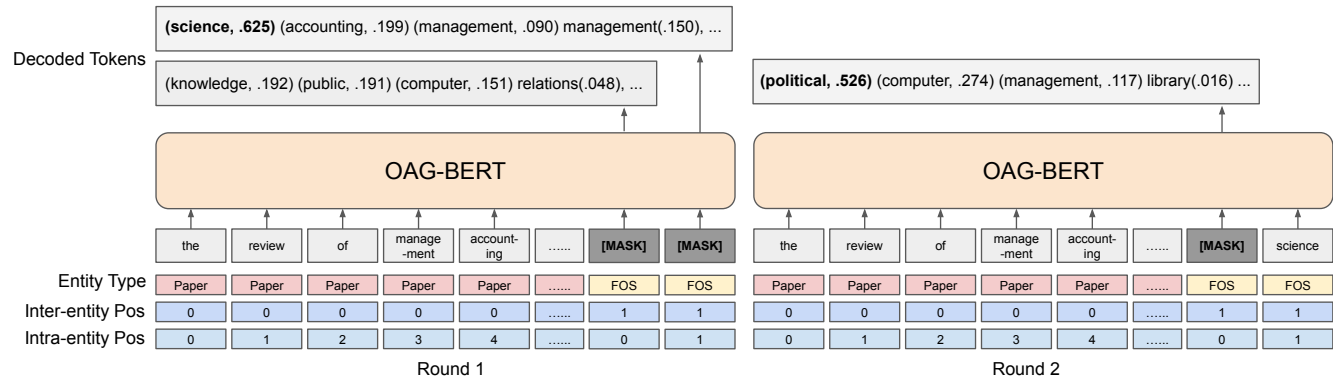


Figure 3: The decoding process of OAG-BERT. The left figure indicates that OAG-BERT decodes the masked token “science” at the second position with the highest probability (0.625) for the first round. Then it decodes “political” at the first position with highest probability (0.526) for the second round as shown in the right figure.

while decoding from scratch, 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.

We design three zero-shot inference tasks to evaluate the entity generation capability of our proposed OAG-BERT and make comparisons with SciBERT, the current state-of-the-art pre-training model in academic domain.

Field-of-Study (FOS) Inference To evaluate the performance of decoding field-of-study, we adopt the research field prediction task from MAG (Microsoft Academic Graph) [38]. 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 predict which research field each paper belongs to.

For each paper, we estimate the probabilities for all FOS candidates and choose the top one. When estimating each one, we concatenate the FOS candidate with the paper title as model input and mask the FOS candidate. For example, when estimating the probability of “computer science”, we add two “[MASK]” tokens to the end of the original title as the input. For OAG-BERT, we treat the newly added “[MASK]” tokens as a new entity, reset entity positions and use FOS entity type embedding additionally. Then we use Equation 2 to calculate the probability. This is denoted as the *Plain* method, as depicted in Figure 3.

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.

Venue and Affiliation Inference Similar to the FOS inference task, we create venue and affiliation inference tasks. From non-pretrained papers, we choose 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. We report the Hit@1 and MRR scores in Table 2.

Results Analysis In Table 2, 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.

表1：所有任务的模型性能摘要。我们仅使用纸张标题作为仅限标题中的输入以及使用其他功能（如FOS或地点）中的最佳性能作为混合中的输入。

Method		1			2			4		
		FOS场地隶属			FOS场地隶属			纸 – 田纸场		
SciBERT	<i>title-only</i>	29.59%	10.03%	8.00%	55.13%	61.86%	35.44%	0.3690	0.4740	0.4570
	<i>mixed</i>	35.33%	18.00%	12.40%	55.63%	78.05%	56.04%	0.4101	-	-
OAG-BERT	<i>title-only</i>	37.33%	22.67%	11.77%	54.54%	63.03%	35.04%	0.4120	0.4892	0.4844
	<i>mixed</i>	49.59%	39.00%	21.67%	64.22%	78.47%	57.63%	0.4823	-	-

— 报告了1,2次命中@ 1用于零拍摄推断和监督分类。3

na是名称歧义的简短。报告了宏成对F1分数。4对于链接预测任务，我们使用预先训练的模型来编码所有类型的节点。

仅为纸节点提供标题。报告了NDCG。

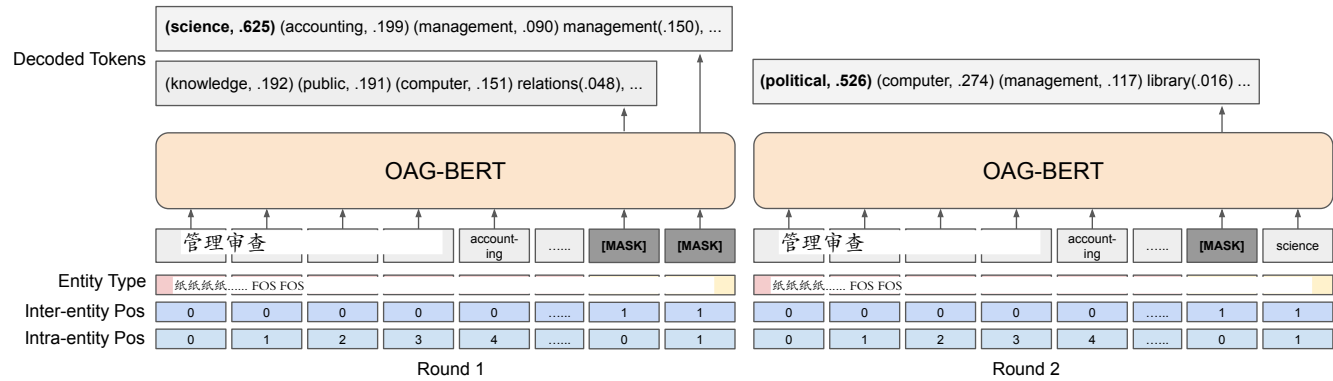


图3：OAG-BERT的解码过程。左图表明OAG-BERT在第二轮的最高概率（0.625）的第二个位置的屏蔽令牌“科学”中解码。然后它在最高概率（0.526）的第一个位置对第二轮解码的“政治”，如右图所示。

在从头开始解码时，我们在预定范围的范围内遍历所有实体长度，具体取决于实体类型，并根据等式2中的计算概率选择顶部候选。

我们设计了三个零拍摄推理任务来评估实体我们拟议的OAG-BERT的一代能力，并制作与SciBERT，目前在学术领域的最先进的培训模型中的融合。

研究视野（FOS）推理，评估解码研究的性能，我们采用MAG（Microsoft学术图）的研究现场预测任务[38]。首先，我们选择19层级的研究领域（FOS），如“生物学”和“计算机科学”。然后，从未在预培训过程中使用的纸质数据，我们随机选择每个FOS的1,000篇论文。任务是预测每张纸张属于哪个研究领域。

对于每篇论文，我们估计所有FOS的概率 – 做了并选择了顶级。当估计每个时，我们将文件候选人与纸张标题作为模型输入和掩盖FOS候选者连接。例如，在估计“计算机科学”的概率时，我们将两个 “[屏蔽]” 令牌添加到原始标题的末尾作为输入。对于OAG-BERT，我们将新添加的 “[屏蔽]” 令牌视为新的实体，重置实体位置并使用FOS实体类型嵌入嵌入。然后我们使用等式2来计算概率。这表示为普通方法，如图3所示。

我们还应用两种技术来改善模型解码表现。第一种技术是向额外的提示单词添加

纸质标题的末尾（在蒙面令牌之前）。我们选择“研究领域：”作为FOS推理任务中的迅速词。第二种技术是将纸张摘要连接到纸张标题的末尾。

场地和隶属关系推断类似于FOS推理任务，我们创建了场地和隶属关系推理任务。从非预用的论文中，我们选择30个最常见的Arxiv类别和30个隶属关系，作为推理候选人，每位候选人都随机选择了100篇论文。包括FOS候选人在内的候选人的全部列表括在附录中。

实验设置完全遵循FOS推理任务，除了我们使用“期刊或地点：”和“附属机构：”分别为迅速言语。OAG-BERT中蒙版实体的实体类型嵌入式也由地点和附属实体类型嵌入式置换相应。我们在表2中报告了HIT @ 1和MRR分数。

结果分析在表2中，我们可以看到建议的Aug-Mented OAG-BERT以大边缘占Scibert。虽然Scibert没有用实体知识预先接受训练，但它仍然比随机猜测更大，这意味着推理任务与纸质内容信息不合适。我们推测纸张内容的预培训过程（如Scibert中使用）还有助于该模型了解其他类型的信息的一般知识，例如研究场外名称或场地名称。

Table 2: The results for zero-shot inference tasks.						
Method	FOS		Venue		Affiliation	
	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR
SciBERT	19.93%	0.37	9.87%	0.22	6.93%	0.19
<i>+prompt</i>	29.59%	0.47	10.03%	0.21	8.00%	0.20
<i>+abstract</i>	25.66%	0.43	18.00%	0.32	10.33%	0.22
<i>+both</i>	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
<i>+prompt</i>	37.33%	0.55	22.67%	0.39	11.77%	0.25
<i>+abstract</i>	49.59%	0.67	39.00%	0.57	21.67%	0.38
<i>+both</i>	49.51%	0.67	38.47%	0.57	21.53%	0.38

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

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

Table 3: The generated FOS for the paper of GPT-3. The gold FOS are bolded. FOS not in the original OAG FOS candidate list 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

Case Study To exhibit the capability of decoding entities, we applied the method described above on the task of FOS generation. Given the paper title and abstract, we use 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 3. In Table 3, the gold FOS are all in the top 16. Some fine-grained entities, though not in candidates, are also generated, such as *Autoregressive language model* or *Few shot learning*.

However, we can still observe some ill-formed or inappropriate entities such as *Architecture* or *Artificial language processing*. While the paper is related to *Model architecture*, the single-token *Architecture* usually refers to the science of designing buildings, which is not suitable in this case. *Artificial language model*, on the other hand, is more like a combination of *Artificial Intelligence* and *Language model*, which have already been generated.

Table 4: The results of the classification task.				
Tasks	Freeze		Finetune	
	SciBERT	OAG-BERT	SciBERT	OAG-BERT
FOS				
<i>title only</i>	33.25 ^{0.25}	43.28 ^{0.12}	55.13 ^{0.30}	54.54 ^{0.29}
<i>+author</i>	30.15 ^{0.07}	41.87 ^{0.06}	55.63 ^{0.42}	55.30 ^{0.43}
<i>+venue</i>	34.77 ^{0.17}	46.99 ^{0.15}	63.18 ^{0.18}	63.53 ^{0.08}
<i>+aff</i>	32.83 ^{0.13}	43.07 ^{0.11}	55.06 ^{0.21}	54.65 ^{0.38}
<i>+all</i>	32.83 ^{0.08}	45.47 ^{0.16}	63.43 ^{0.15}	64.22 ^{0.38}
Venue				
<i>title only</i>	24.62 ^{0.52}	32.87 ^{1.47}	61.86 ^{0.32}	63.03 ^{0.46}
<i>+author</i>	21.21 ^{0.82}	30.91 ^{0.96}	62.62 ^{0.34}	63.46 ^{0.48}
<i>+aff</i>	24.38 ^{0.49}	32.32 ^{1.36}	62.13 ^{0.43}	62.65 ^{0.49}
<i>+fos</i>	40.49 ^{1.25}	52.61 ^{0.79}	78.05 ^{0.14}	78.47 ^{0.25}
<i>+all</i>	39.92 ^{1.17}	51.33 ^{0.44}	77.88 ^{0.16}	78.34 ^{0.62}
Affiliation				
<i>title only</i>	13.88 ^{0.83}	19.72 ^{0.64}	35.44 ^{0.45}	35.04 ^{0.61}
<i>+author</i>	20.65 ^{1.04}	32.19 ^{0.92}	52.68 ^{0.18}	53.33 ^{0.43}
<i>+venue</i>	16.57 ^{0.60}	25.23 ^{0.72}	43.13 ^{0.36}	43.65 ^{0.40}
<i>+fos</i>	17.39 ^{0.86}	22.06 ^{0.37}	37.05 ^{0.80}	37.60 ^{0.51}
<i>+all</i>	24.02 ^{0.87}	32.49 ^{0.50}	56.04 ^{0.95}	57.63 ^{0.49}

In summary, although our proposed OAG-BERT model is not born for decoding, it still exhibits the potential of generating high-quality entities in the zero-shot settings.

4.2 Supervised Classification

In this section, we develop the supervised classification tasks on top of the datasets described above, which are enlarged by 10 times following the same generating process. The data in the zero-shot inference are kept as test sets. We construct validation sets to select the best models during fine-tuning, with the same size as the test sets. The rest data are used as training sets. The sizes of all datasets for all tasks are enclosed in the appendix.

In supervised classification tasks, we remove the masked tokens and feed the averaged output embeddings from the pre-training models to a single fully-connected layer. We apply softmax layer to make predictions at last. As for the inputs, to present the effectiveness of heterogeneous entity types, we not only use paper titles as inputs but also concatenate other entities. Besides, we also tested the model performance with and without the original pre-training model parameters frozen. We follow the standard configurations for fine-tuning BERT, which are enclosed in the appendix.

As shown in Table 4, the OAG-BERT outperforms SciBERT by a large margin when the parameters in pre-trained parts are frozen. When not frozen, for venue and affiliation prediction, OAG-BERT surpasses SciBERT significantly. In FOS prediction, although OAG-BERT under-performs SciBERT in some cases, the best performance for using all available entities in OAG-BERT still beats the one reached by SciBERT.

We also observe that different types of entities contribute to various tasks in dissimilar ways. For example, the use of author information is particularly helpful for affiliation prediction but not very useful in FOS prediction. On the other hand, the field of study (FOS) inputs, work pretty well in venue prediction but provide marginal improvements to affiliation prediction.

表2：零拍摄推理任务的结果。						
Method	FOS		Venue		Affiliation	
	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR
SciBERT	19.93%	0.37	9.87%	0.22	6.93%	0.19
<i>+prompt</i>	29.59%	0.47	10.03%	0.21	8.00%	0.20
<i>+abstract</i>	25.66%	0.43	18.00%	0.32	10.33%	0.22
<i>+both</i>	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
<i>+prompt</i>	37.33%	0.55	22.67%	0.39	11.77%	0.25
<i>+abstract</i>	49.59%	0.67	39.00%	0.57	21.67%	0.38
<i>+both</i>	49.51%	0.67	38.47%	0.57	21.53%	0.38

我们还观察到拟议的摘要可以随时使用有助于提高性能。另一方面，迅速单词与Scibert运行良好，但仅对OAG-BERT提供有限的帮助。此外，隶属关系似乎比其他两个任务更难。A.1提供了进一步的分析。还封闭了两个扩展实验，揭示了两种结果：

- （1）使用令牌日志概率的求和作为实体日志概率比使用平均值更好。
- （2）无序解码更适合于基于编码器与左右解码相比，Scibert和OAG-BERT这样的模型。

表3：GPT-3纸产生的FOS。金fos大胆。未在原始OAG FOS候选列表中的FOS下划线。

标题语言模型很少拍摄学习者	
摘要最近的工作通过预先培训了许多NLP任务和替补标志，通过对特定任务进行微调，展示了许多NLP任务和基层的大量收益。虽然通常在架构中的任务不可知，但这种方法仍然需要千万或成千上万例的任务特定的微调数据集。相比之下，人类通常可以.....	
Generated FOS	自然语言处理，自归语言模型，计算机科学，句子，人工智能，域适应，语言模型，少量射击学习，大型语料库，算术，机器学习，建筑，理论计算机科学，数据挖掘，语言学，
金FOS语言模型，计算机科学，语言学	

案例研究表明解码实体的能力，我们应用了上面描述的方法对FOS生成的任务。鉴于纸张标题和摘要，我们使用宽度为16的波束搜索来解码FOS实体。我们从单令牌实体搜索四边形令牌实体。表3中列出了前16名生成的。在表3中，金FOS都在前16名中。还产生了一些细粒度的实体，但不在候选人中，例如自回归语言模型或少量射击学习。

但是，我们仍然可以观察一些不成熟的或不承认的在架构或人工语言处理等实体。虽然论文与模型架构有关，但单令架构通常是指设计建筑物的科学，这在这种情况下不适合。另一方面，人工语言模型更像是已经生成的人工智能和语言模型的组合。

表4：分类任务的结果。				
Tasks	Freeze		Finetune	
	SciBERT	OAG-BERT	SciBERT	OAG-BERT
FOS				
<i>title only</i>	33.25 ^{0.25}	43.28 ^{0.12}	55.13 ^{0.30}	54.54 ^{0.29}
<i>+author</i>	30.15 ^{0.07}	41.87 ^{0.06}	55.63 ^{0.42}	55.30 ^{0.43}
<i>+venue</i>	34.77 ^{0.17}	46.99 ^{0.15}	63.18 ^{0.18}	63.53 ^{0.08}
<i>+aff</i>	32.83 ^{0.13}	43.07 ^{0.11}	55.06 ^{0.21}	54.65 ^{0.38}
<i>+all</i>	32.83 ^{0.08}	45.47 ^{0.16}	63.43 ^{0.15}	64.22 ^{0.38}
Venue				
<i>title only</i>	24.62 ^{0.52}	32.87 ^{1.47}	61.86 ^{0.32}	63.03 ^{0.46}
<i>+author</i>	21.21 ^{0.82}	30.91 ^{0.96}	62.62 ^{0.34}	63.46 ^{0.48}
<i>+aff</i>	24.38 ^{0.49}	32.32 ^{1.36}	62.13 ^{0.43}	62.65 ^{0.49}
<i>+fos</i>	40.49 ^{1.25}	52.61 ^{0.79}	78.05 ^{0.14}	78.47 ^{0.25}
<i>+all</i>	39.92 ^{1.17}	51.33 ^{0.44}	77.88 ^{0.16}	78.34 ^{0.62}
Affiliation				
<i>title only</i>	13.88 ^{0.83}	19.72 ^{0.64}	35.44 ^{0.45}	35.04 ^{0.61}
<i>+author</i>	20.65 ^{1.04}	32.19 ^{0.92}	52.68 ^{0.18}	53.33 ^{0.43}
<i>+venue</i>	16.57 ^{0.60}	25.23 ^{0.72}	43.13 ^{0.36}	43.65 ^{0.40}
<i>+fos</i>	17.39 ^{0.86}	22.06 ^{0.37}	37.05 ^{0.80}	37.60 ^{0.51}
<i>+all</i>	24.02 ^{0.87}	32.49 ^{0.50}	56.04 ^{0.95}	57.63 ^{0.49}

总之，虽然我们提出的OAG-BERT模型不是出生于解码，它仍然表现出在零拍摄环境中产生高质量实体的可能性。

4.2 Supervised Classification

在本节中，我们在上述数据集的顶部开发了监督分类任务，该任务在同一生成过程之后的10次上放大了10次。零拍摄推断中的数据保存为测试集。我们构建验证集以在微调期间选择最佳模型，与测试集相同的大小。其余数据用作训练集。所有任务的所有数据集的大小都包含在附录中。

在监督分类任务中，我们删除了蒙面标记并将平均输出嵌入从预训练模型中馈送到单个完全连接的层。我们应用Softmax层以终止进行预测。至于输入，要呈现并构实体类型的有效，我们不仅使用纸张标题作为输入，而且还连接其他实体。此外，我们还通过冻结的原始预训练模型参数测试了模型性能。我们遵循封装在附录中的微调BERT的标准配置。

如表4所示，OAG-BERT Outperferflys scibert预训练零件中的参数被冻结时大边距。当不冷冻时，对于场地和隶属关系，OAG-BERT显着超过Scibert。在FOS预测中，虽然OAG-BERT在某些情况下执行SCIBERT，但在某些情况下，使用OAG-BERT中的所有可用实体的最佳性能仍然击败Scibert达到的。

我们还观察到不同类型的实体有助于各种各样的任务以不同的方式。例如，作者信息的使用对于来自FOS预测而不是非常有用的隶属关系。另一方面，研究领域（FOS）输入，在地点预测中工作得很好，但为隶属关系提供了边际改善。

In conclusion, the proposed OAG-BERT is effective in both zero-shot tasks and supervised tasks. The additionally learned heterogeneous entities can help the model reach better performance while dealing with multiple types of inputs.

Table 5: The Macro Pairewise F1 scores for the name disambiguation task.

Inputs	SciBERT	OAG-BERT
<i>title</i>	0.3690	0.4120
<i>+fos</i>	0.4101	0.4643
<i>+venue</i>	0.3603	0.4247
<i>+fos+venue</i>	0.3903	0.4823
Leader Board Top 1		0.4900

4.3 Name Disambiguation

Previous experiments focus on the decoding capability and fine-tuning performance on downstream tasks. In this section, we adopt the name disambiguation problem to validate the paper representation quality produced by OAG-BERT. In this problem, given a set of papers with authors of the same name, the designed algorithm needs to separate these papers into several 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*³ [3, 49] and apply the embeddings generated by pre-trained models to solve name disambiguation from scratch. Formally, for each paper, we use the paper title and other attributes such as field-of-study or published venue as model input. Then we average over all the output token embeddings for the paper and use it as the paper embedding. After that, 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. The threshold from the best validation results is used on test set evaluation. We calculated the macro pairwise f1 score following previous works.

The results in Table 5 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 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 5 we only list the results with title, field-of-study, and venue as inputs. 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 leaderboard⁴ and find that our proposed OAG-BERT reaches close performance as compared with the top 1.

4.4 Link Prediction

In previous sections, we present the effectiveness of using OAG-BERT individually. In this section, we apply the heterogeneous entity embeddings of OAG-BERT as pre-trained initializations for node 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] and combine it with the pre-trained embeddings from OAG-BERT.

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 [44] to encode titles as input features. For other types of nodes, HGT use metapath2vec [10] to initialize the features.

However, there are two problems with using XLNet on the heterogeneous academic graph. First, the XLNet was pre-trained on universal language corpus, which is lack of academic domain data. Second, XLNet can only encode paper nodes by using their titles and is unable to generate useful embeddings for other types of nodes like author or affiliation.

To this end, we propose to 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 prove 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 HGT⁵. 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 7. 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.

4.5 NLP Tasks

Previous experiments have demonstrated the superiority of OAG-BERT on tasks involving multi-type entities. In this section, we will further explore the performance of OAG-BERT on natural language processing tasks, which only contain text-based information such as paper titles and abstracts. We will show that although pre-trained with heterogeneous entities, the OAG-BERT can still perform competitive results with SciBERT on NLP tasks.

We made comparisons over three models, including **SciBERT** (both the original paper results and the reproduced results), **S2ORC** (similar to SciBERT except pre-trained with more data), and **OAG-BERT** (both the vanilla version and the augmented version).

In accord with SciBERT [2], we evaluate the model performance on the same 12 NLP tasks, including Named Entity Recognition (NER), Dependency Parsing (DEP), Relation Extraction (REL), PICO Extraction (PICO), and Text Classification (CLS). These tasks only focus on single sentence representation so we add another three sequential sentence classification (SSC) tasks used in [5], to further verify the capability of pre-training models on long texts. The evaluation metrics are also accord with the usage in SciBERT [2] and

总之，所提出的OAG-BERT在零中有效拍摄任务和监督任务。另外学习的异构实体可以帮助模型达到更好的性能，同时处理多种类型的输入。

表5：名称消除歧义任务的宏成对F1分数。

Inputs	SciBERT	OAG-BERT
<i>title</i>	0.3690	0.4120
<i>+fos</i>	0.4101	0.4643
<i>+venue</i>	0.3603	0.4247
<i>+fos+venue</i>	0.3903	0.4823
Leader Board Top 1		0.4900

4.3 Name Disambiguation

以前的实验专注于下游任务上的解码能力和微调性能。在本节中，我们采用名称消歧问题来验证OAG-BERT生产的纸张代表质量。在这个问题中，给出了一组与同名作者的论文，所设计的算法需要将这些论文分成几个集群，同一群集中属于同一作者，不同的集群代表不同的作者。

我们使用公共数据集WhoisWho-V13 [3,49]并应用

由预先训练的模型生成的嵌入，以解决从头划分的名称。正式地，对于每份纸张，我们使用纸质标题和其他属性，例如学习场或公布的场地作为模型输入。然后我们平均过往纸张的所有输出令牌嵌入，并将其用作纸张嵌入。之后，我们使用所有文件构建一个图形作为图形节点，并设置一个阈值以选择边缘。边缘在纸张之间，它们嵌入的成对余弦相似性大于阈值。最后，对于图表中的每个连接的组件，我们将其视为群集。在验证集中搜索从0.65到0.95的阈值。来自最佳验证结果的阈值用于测试集评估。我们计算了以前的作品后宏成对F1分数。

表5中的结果表明OAG-BERT的嵌入

比在作者名称歧义中直接使用的同时明显优于斯科德嵌入。我们还观察到，对于Scibert，最好的阈值始终为0.8，而OAG-BERT的该值是0.9，这反映了由OAG-BERT产生的纸张嵌入通常比Scibert生产的纸张嵌入式更近。

在表5中，我们只列出了标题，研究和研究的结果地点作为投入。虽然我们试图使用摘要，作者和隶属关系，但没有预期的性能改善。我们推测它是因为这些类型的信息更复杂，可能需要额外的分类器头或微调，作为上述监督分类任务。此外，我们还报告了名称歧义挑战队排行榜4中的前1名得分，并发现我们提出的OAG-BERT与前1名相比达到了密切的性能。

4.4 Link Prediction

在前面的部分中，我们介绍了单独使用OAG-BERT的有效性。在本节中，我们将OAG-BERT的异构实体嵌入作为学术图中的节点嵌入的预先训练的初始化，并显示OAG-BERT也可以与其他类型的模型一起使用。具体而言，我们从[16]中获取异构图形变压器（HGT）模型，并将其与来自OAG-BERT的预先培训的嵌入式组合。

为了使异构图中的链接预测，

HGT
First提取节点功能的作者，然后将HGT层应用于编码图形功能。对于纸张节点，作者使用XLnet
[44]将标题进行编码为输入功能。对于其他类型的节点，HGT使用Metapath2VEC [10]初始化功能。

但是，在HET上使用XLNET存在两个问题 – 不均匀的学术图。首先，XLNET在普通语言语料库上预先培训，这是缺乏学术域数据。其次，XLNET只能使用其标题编码纸质节点，并且无法为其他类型的节点生成有用的嵌入品，如作者或隶属度。

为此，我们建议替换原始的XLNET编码器

通过我们的OAG-BERT模型，可以解决上述两个挑战。我们使用OAG-BERT模型来编码所有类型的节点并使用生成的嵌入式作为其节点功能。为了证明OAG-BERT对编码异质节点的有效性，我们还比较Scibert与OAG-BERT的性能。我们在HGT5发布的CS数据集上进行了实验。数据集的详细信息在附录中传递。

纸张领域和纸张的NDCG和MRR分数

表7中报告了链路预测。它表明，由于大型科学语料库的预训练，Scibert显著地通过了原始的XLNET性能。我们提出的OAG-BERT进一步改进，因为它可以更好地了解异构图中的实体知识。

4.5 NLP Tasks

以前的实验已经证明了OAG-BERT对涉及多型实体的任务的优越性。在本节中，我们将进一步探讨OAG-BERT对自然语言处理任务的性能，只包含基于文本的信息，如纸张标题和摘要。我们将表明，虽然具有异构实体预先训练，但OAG-BERT仍然可以使用SCIBERT对NLP任务进行竞争结果。

我们通过三种模型进行了比较，包括斯科尔特（原始纸质结果和再现结果），S2ORC（类似于除了使用更多数据预培训的斯科尔特）和OAG-BERT（Vanilla版本和增强版）。

根据Scibert [2]，我们评估模型性能

在相同的12个NLP任务中，包括命名实体识别（ner），依赖解析（dep），关系提取（rel），pico提取（pico）和文本分类（cls）。这些任务仅关注单句表示，因此我们添加了[5]中使用的另外三个顺序句子分类（SSC）任务，以进一步验证长文本的预培训模型的能力。评估指标也符合SCIBERT [2]的用途

³<https://www.aminer.cn/whoiswho>

⁴<https://www.biendata.xyz/competition/aminer2019/leaderboard/>

⁵<https://github.com/acbull/pyHGT>

³<https://www.aminer.cn/whoiswho>

⁴<https://www.biendata.xyz/competition/aminer2019/leaderboard/>

⁵<https://github.com/acbull/pyHGT>

Table 6: The results for NLP Tasks.

Field	Task	Dataset	Samples ¹	S2ORC	SciBERT		OAG-BERT	
					Reported ²	Reproduced	Vanilla	Augmented
Bio	NER	BC5CDR [26]	3942	90.04 ^{0.06}	90.01	89.77 ^{.23}	89.71 ^{.13}	89.71 ^{.12}
		JNLPBA [21]	16807	77.70 ^{.25}	77.28	77.29 ^{.38}	75.81 ^{.20}	76.99 ^{.04}
		NCBI-disease [9]	5424	88.70 ^{.52}	88.57	88.10 ^{.06}	87.90 ^{.12}	88.77 ^{.56}
	PICO	EBM-NLP [33]	27879	72.35 ^{.95}	72.28	72.52 ^{.71}	72.22 ^{.24}	71.74 ^{.49}
	DEP	GENIA - LAS [20]	14326	90.80 ^{.19}	90.43	90.57 ^{.08}	89.99 ^{.10}	90.12 ^{.11}
		GENIA - UAS [20]		92.31 ^{.18}	91.99	92.12 ^{.07}	91.57 ^{.08}	91.63 ^{.09}
	REL	ChemProt [24]	4169	84.59 ^{.93}	83.64	83.46 ^{.28}	82.14 ^{1.12}	80.21 ^{1.42}
	SSC	Pubmed-RCT-20k [7]	15130	-	92.90	92.86 ^{.12}	92.80 ^{.05}	92.73 ^{.08}
CS		NICTA-piboso [22]	735	-	84.80	83.93 ^{.58}	83.02 ^{.67}	84.00 ^{.32}
	NER	SciERC [31]	1861	68.93 ^{.19}	67.57	66.28 ^{.20}	67.80 ^{.24}	66.75 ^{.77}
	REL	SciERC [31]	3219	81.77 ^{1.64}	79.97	80.21 ^{.88}	76.59 ^{0.75}	78.63 ^{.06}
	CLS	ACL-ARC [18]	1688	68.45 ^{2.47}	70.98	70.34 ^{3.07}	66.13 ^{1.58}	64.79 ^{3.35}
	SSC	CSAbstract [5]	1668	-	83.10	82.40 ^{.33}	82.48 ^{.44}	82.59 ^{.67}
Multi	CLS	Paper Field [38]	84000	65.99 ^{.08}	65.71	65.77 ^{.13}	64.67 ^{.14}	64.95 ^{.10}
		SciCite [4]	7320	84.76 ^{.37}	85.49	85.65 ^{.54}	85.25 ^{.38}	84.95 ^{.32}

¹ *Samples* refers to the number of training samples in the dataset.

² We run the fine-tuning process for 5 times with different random seeds and report the mean and standard deviation. The results in the original paper of SciBERT do not report this. The results for NER tasks in the original SciBERT model use a different casing version of pre-trained model while all other results are achieved by uncased pre-trained models.

Table 7: The result of link prediction tasks.

Tasks	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

Sequential-Sentence-Classification [5], which can be found in the appendix along with the task details and hyper-parameter settings.

The results in Table 6 show that the proposed OAG-BERT is competitive with SciBERT and a bit behind the S2ORC. Comparing with the reproduced SciBERT, our vanilla OAG-BERT only shows clear disadvantages on the SciERC REL task and the ACL-ARC CLS task, where datasets are relatively small and are sensitive to a few swinging samples. We ascribe the minor differences in other tasks to the differences in training corpus and the data cleaning techniques. The augmented OAG-BERT, although trained with heterogeneous entities that differ from the inputs of downstream NLP tasks, still presents similar performance to the vanilla version.

In summary, despite the fact that the OAG-BERT does not surpass the previous state-of-the-art academic pre-training model on NLP tasks, it still keeps the knowledge on these language dedicated tasks even after pre-training with multiple types of entities.

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 reviewer recommendation problem.

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 [40] 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.

Moreover, we release the OAG-BERT model in CogDL package, helping users take advantages of our OAG-BERT model in their own applications.

6 CONCLUSION

In conclusion, we propose a new pre-training model in the academic domain, called OAG-BERT. Compared to previous models like SciBERT, the OAG-BERT incorporates entity knowledge during pre-training, which benefits lots of downstream tasks that involve multi-type entities, such as name disambiguation or link prediction on the heterogeneous academic graph. We apply OAG-BERT to real-world applications, which improves the efficiency of these applications. We finally release the pre-trained model in CogDL, providing free use to arbitrary users.

There are still some problems remained. First, although OAG-BERT can decode entities, it is hard to generate long entities efficiently, due to the exhaustive search for the entity length. Second, the learning for sparse entities such as author names is much less effective than other entities due to the lack of pre-training, which

表6：NLP任务的结果。

Field	Task	Dataset	Samples ¹	S2ORC	SciBERT		OAG-BERT	
					Reported ²	Reproduced	Vanilla	Augmented
Bio	NER	BC5CDR [26]	3942	90.04 ^{0.06}	90.01	89.77 ^{.23}	89.71 ^{.13}	89.71 ^{.12}
		JNLPBA [21]	16807	77.70 ^{.25}	77.28	77.29 ^{.38}	75.81 ^{.20}	76.99 ^{.04}
		NCBI-disease [9]	5424	88.70 ^{.52}	88.57	88.10 ^{.06}	87.90 ^{.12}	88.77 ^{.56}
	PICO	EBM-NLP [33]	27879	72.35 ^{.95}	72.28	72.52 ^{.71}	72.22 ^{.24}	71.74 ^{.49}
	DEP	GENIA - LAS [20]	14326	90.80 ^{.19}	90.43	90.57 ^{.08}	89.99 ^{.10}	90.12 ^{.11}
		GENIA - UAS [20]		92.31 ^{.18}	91.99	92.12 ^{.07}	91.57 ^{.08}	91.63 ^{.09}
	REL	ChemProt [24]	4169	84.59 ^{.93}	83.64	83.46 ^{.28}	82.14 ^{1.12}	80.21 ^{1.42}
	SSC	Pubmed-RCT-20k [7]	15130	-	92.90	92.86 ^{.12}	92.80 ^{.05}	92.73 ^{.08}
CS		NICTA-piboso [22]	735	-	84.80	83.93 ^{.58}	83.02 ^{.67}	84.00 ^{.32}
	NER	SciERC [31]	1861	68.93 ^{.19}	67.57	66.28 ^{.20}	67.80 ^{.24}	66.75 ^{.77}
	REL	SciERC [31]	3219	81.77 ^{1.64}	79.97	80.21 ^{.88}	76.59 ^{0.75}	78.63 ^{.06}
	CLS	ACL-ARC [18]	1688	68.45 ^{2.47}	70.98	70.34 ^{3.07}	66.13 ^{1.58}	64.79 ^{3.35}
	SSC	CSAbstract [5]	1668	-	83.10	82.40 ^{.33}	82.48 ^{.44}	82.59 ^{.67}
Multi	CLS	Paper Field [38]	84000	65.99 ^{.08}	65.71	65.77 ^{.13}	64.67 ^{.14}	64.95 ^{.10}
		SciCite [4]	7320	84.76 ^{.37}	85.49	85.65 ^{.54}	85.25 ^{.38}	84.95 ^{.32}

1个样本是指数据集集中的训练样本的数量。2我们用不同的随机种子运行微调过程5次，并报告平均值和标准偏差。这

结果在斯科尔特的原文中没有报告这一点。原始SCIBERT模型中的NER任务的结果使用预先训练模型的不同套管版本，而所有其他结果通过未应用的预先训练的模型实现。

表7：链路预测任务的结果。

Tasks	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

顺序句子分类[5]，可以在附录中以及任务详细信息和超参数设置中找到。

表6中的结果表明，所提出的OAG-BERT是竞争Scibert和S2ORC后面有点竞争。与转载的Scibert相比，我们的Vanilla OAG-BERT仅显示了Scienc Rel任务和ACL-ARC CLS任务的明确缺点，其中数据集相对较小，对几个摆动样本敏感。我们将其他任务中归类对培训语料库和数据清洁技术的差异进行了次要差异。增强的OAG-BERT，虽然具有与下游NLP任务的输入不同的异构实体，但仍然对香草版具有类似的性能。

总之，尽管OAG-BERT没有超过的事实即使在使用多种类型的实体预训练之后，它仍然在NLP任务上的先前最先进的学术预培训模型，即使在预训练之后，它仍然会对这些语言专用任务进行了解。

5 DEPLOYED APPLICATIONS

在本节中，我们将介绍我们的OAG-BERT的几个现实世界应用程序。

首先，歧义任务的结果表明OAG-BERT在使用多型实体编码纸质信息时相对强大，这进一步帮助为纸作者产生代表嵌入式。因此，我们将OAG-BERT应用于审稿人的建议问题。

为了解决这个问题，我们与阿里巴巴合作并发展在OAG-BERT顶部的一种实用算法，可以自动为应用程序分配适当的审稿人，并大大利益审查过程。

除此之外，我们还将OAG-BERT融为一体 – AMiner [40]系统的陈大果部件。在AMiner，我们利用OAG-BERT处理有关学术性的最丰富的信息。例如，随着解码FOS实体的能力，我们使用OAG-BERT自动生成FOS，用于未标记的文件。此外，我们将OAG-BERT与名称歧义框架相似。最后，我们使用OAG-BERT为用户推荐相关论文，利用其在编码纸张嵌入方面的能力。

此外，我们在COGDL包中释放OAG-BERT模型，帮助用户在自己的应用程序中采取OAG-BERT模型的优势。

6 CONCLUSION

总之，我们提出了一种新的培训模型，称为OAG-BERT。与以前的模型如Scibert这样的模型，OAG-BERT在预培训期间结合了实体知识，这有利于许多涉及多型实体的下游任务，例如在异构学术图上的名称消歧或链路困扰。我们将OAG-BERT应用于真实世界的应用，这提高了这些应用的效率。我们终于在COGDL中释放了预先训练的模型，为任意用户提供免费使用。

仍然存在一些问题。首先，虽然oag-BERT可以解码实体，由于实体长度的详尽搜索，难以产生长实体。其次，由于缺乏预先训练，诸如作者名称等稀疏实体的学习远低于其他实体，这

hinders the downstream tasks to fully leverage the entity information. We leave these problems for future explorations.

REFERENCES

- [1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262* (2018).
- [2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).
- [3] 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).
- [4] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608* (2019).
- [5] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054* (2019).
- [6] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).
- [7] Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071* (2017).
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [9] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47 (2014).
- [10] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*.
- [11] Zhengxiao Du, Jie Tang, and Yuhui Ding. 2018. Polar: Attention-based cnn for one-shot personalized article recommendation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.
- [12] Zhengxiao Du, Jie Tang, and Yuhui Ding. 2019. POLAR++: Active One-shot Personalized Article Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2019).
- [13] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*.
- [14] 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).
- [15] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. 2020. Gpt-gnn: Generative pre-training of graph neural networks. In *SIGKDD*.
- [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] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *TACL* 6 (2018).
- [19] Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. 2019. A scalable hybrid research paper recommender system for microsoft academic. In *WWW*.
- [20] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics* 19, suppl_1 (2003).
- [21] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *JNLPBA*. Citeseer.
- [22] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, Vol. 12. Springer.
- [23] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [24] Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. ChemProt-3.0: a global chemical biology diseases mapping. *Database* 2016 (2016).
- [25] 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).
- [26] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016 (2016).

- [27] 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.
- [28] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. 2020. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218* 1, 2 (2020).
- [29] 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).
- [30] 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).
- [31] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602* (2018).
- [32] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacey: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669* (2019).
- [33] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*, Vol. 2018. NIH Public Access.
- [34] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).
- [35] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. E-bert: Efficient-yet-effective entity embeddings for bert. *arXiv preprint arXiv:1911.03681* (2019).
- [36] 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).
- [37] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deep-speed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*.
- [38] 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*.
- [39] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized Language and Knowledge Embedding. *arXiv preprint arXiv:2010.00309* (2020).
- [40] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-miner: extraction and mining of academic social networks. In *SIGKDD*.
- [41] 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.
- [42] 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).
- [43] Kan Wu, Jie Tang, and Chenhui Zhang. 2018. Where Have You Been? Inferring Career Trajectory from Academic Social Network.. In *IJCAI*.
- [44] 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).
- [45] Yan Wang Qibin Chen Yizhen Luo Xingcheng Yao Aohan Zeng Shiguang Guo Peng Zhang Guohao Dai Yu Wang Chang Zhou Hongxia Yang Jie Tang Yukuo Cen, Zhenyu Hou. 2021. CogDL: An Extensive Toolkit for Deep Learning on Graphs. *arXiv preprint arXiv:2103.00959* (2021).
- [46] 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*.
- [47] Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. ProNE: Fast and Scalable Network Representation Learning.. In *IJCAI*, Vol. 19.
- [48] Minjia Zhang and Yuxiong He. 2020. Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping. *arXiv preprint arXiv:2010.13369* (2020).
- [49] Yutao Zhang, Fanjin Zhang, Peiran Yao, and Jie Tang. 2018. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.. In *SIGKDD*.
- [50] 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).

阻碍下游任务以充分利用实体信息。我们为未来的探索留下了这些问题。

REFERENCES

- [1] Walid Ammar, Dirk Growelld, Moon Bhagwatal等。2018年。建造语义学者文学图中的文献图。Arxiv预印迹arxiv: 1805.02262（2018）。
- [2] Iz Beltagy, Kyle Lo和Arman Cohan。2019。Scibert：一种预用的语言科学文本的模型。Arxiv预印迹Arxiv: 1903.10676（2019）。
- [3] BO Chen, jin钊Zhang, J IE tang, Lin规范CAI, Zhao欲Wang, Shu Zhao, hong陈, 捏着李。2020.
- [4] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen和Field Cady。2019年。科学出版物的引文意图分类的结构脚手架。Arxiv预印迹arxiv: 19
- [5] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi和Daniel S焊接。2019年。序列句子分类的预付费语言模型。Arxiv预印迹arxiv: 1909.04054
- [6] Z i行D爱, zh i临yang, Y i明yang, Jaime carbon El蓝, quo CV LE, Andrus蓝Salakhutdinov。2019.转换器–XL：超薄语言模型超出固定长度上
- [7] Franck Dernoncourt和Ji Young Lee。2017。PubMed 200k RCT：序列的数据集 – 医疗摘要中的Tial句子分类。Arxiv预印迹arxiv: 1710.06071（201
- [8] Jacob Devlin, Ming–Wei Chang, Kent on Lee, and Kristina to U譚OVA。2018。Bert：用于语言理解的深双向变压器的预培训。Arxiv预印迹arxiv: 1
- [9] Rezarta Islamaj Do an, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease 语料库：疾病名称识别和概念标准化的资源。中国生物医学信息学杂志47（2014）。
- [10]
- [11] zh Eng小D U, J IE tang, Andy U惠ding. 2018. polar: attention–based CNN for 一次性个性化文章推荐。在数据库中联合欧洲机器学习会议和知
- [12] zh Eng小D U, J IE tang, Andy U惠ding. 2019. polar++: active one–shot 个性化文章推荐。IEEE知识和数据工程交易（2019）。
- [13] Aditya Grover和Jure Leskovec。2016. node2vec: 可扩展功能学习SIGKDD
- [14] Doug Downey, 诺亚史密斯。2020. 不要停止预训练：将语言模型适应域和任务。Arxiv预印迹arxiv日期: 2004.10964（2020）。
- [15] Z i钊hu, Y u小dong, Ku按San Wang, Kai–Wei Chang, Andy i周sun. 2020. GPT–GNN：Graph神经网络的生成预训练。在SIGKDD。
- [16] Z i钊hu, Y u小dong, Ku按San Wang, Andy i周sun. 2020. heterogeneous 图形变压器。在www。
- [17] man达人Jo是, Dan起Chen, yin蔡I IU, Daniels weld, Luke Z哦TT了摸鱼儿, 安定omer levy。2020. Spanbert：通过代表和预测跨度来改善预训练。TACL 8（2020）。
- [18] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland和Dan Jurafsky。2018年通过引文框架测量科学领域的演变。TACL 6（2018）。
- [19] Ans hulk Ana KIA, Z Hi红S很, Darrin EI的, and Ku按San Wang. 2019. A scalable
- “20” J–D Kim, Tomoko Yuka, Yuka Toshi, Achi'Ichi'Ichiji。2003年。kao nip s–a 用于生物教学的语义注释语料库。生物信息学19, SUPPL_1（20
- “21” Junkon G MI, Tomoko H, Yoshimasa Tsuru, Yuka Tatsu, D体L钥匙R. 2004.
- [22]是Nam Kim, David Martinez, Lawrence Cavedon和Lars Yencken。2011年基于证据的句子的最新分类。在BMC生物信息学中, Vol. 12.弹
- [23]托马斯n kipf和最大的好处。2016年, 半监督分类与图形
- [24] Jens Kringelum, Sonny Kringelum, S renBrunak, Ole Lund, Oprea的Tudor, 和橄榄球。2016. ChemProt–3.0: 全球化学生物学疾病测绘。数据库2016（2016）。
- [25] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, 和Jaewoo Kang。2020。Biobert: 生物医学文本挖掘的预先训练的生物医学语言表示模型。生物信息学36,4（2020）。
- [26] J i Aoli, Y UE屏sun, Robin J Johnson, Daniel as CIA可以, x H–H酸Wei, Robert 李曼, 艾伦彼得戴维斯, 卡罗琳·杰斯利, 托马斯C Wiegers和Zhiyong Lu。2016年。生物重建v CDR任务语料库：化学疾病关系提取资源。数据库2016（2016）。

- [27] Wei矛L IU, peng Zhou, zh EZ好, Z Hi若Wang, Q IJ U, H熬汤Deng, and ping王。2020. K–BERT：使用知识图表启用语言表示。在Aaai, Vol. 34。
- [28] 唐。2020.自我监督学习：生成或对比。ARXIV预印符号ARXIV：2006.08218 1,2（2020）。
- [29] Levy, Mike Lewis, Luke Zettlemoyer和Veselin Stoyanov。2019. Roberta：一种稳健优化的伯特预用方法。Arxiv预印迹arxiv: 1907.11692（2019）。
- [30] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney和Dan S Weld。2019年。S2ORC：语义学者开放研究语料库。Arxiv预印迹arxiv: 1911.02
- [31] Y IL u按, I U横he, Mario斯特恩do RF, and Hanna NE和ha级石人子。2018. multi–task 识别科学知识图形建设的实体, 关系和贯穿研究。Arxiv预印迹ar
- [32] Mark Neumann, Daniel King, Iz Beltagy和Waleed Ammar。2019. SCISPACY: 快生物医学自然语言处理的强大模型。Arxiv预印迹Arxiv: 1902.076
- [33] Benjamin N页, Jun以Jess YL i, Roma Patel, yin非yang, I爱NJ Marshall, ani Nenkova和Byron C Wallace。2018.一种具有多级注释的患者, 干预和结果的语料库, 以支持医学文献的语言处理。在ACL, Vol. 2018. NIH公共获取。
- [34] Fabio Petroni, TimRockt schel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller和Sebastian Riedel。2019.语言模型作为知识库吗? Arxiv预印迹arxiv: 1909.01066（2019）。
- [35] 伯特的有效实体嵌入。Arxiv预印迹Arxiv: 1911.03681（2019）。
- [36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei和Ilya Sutskever。2019年。语言模型是无人监督的多任务学习者。Open
- [37] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase和Yuxiong他。2020.深–速度：系统优化使培训具有超过1000亿个参数的深度学习模型。
- [38] AR那边sin哈, Z Hi红S很, yang song, ha OMA, Darrin EI的, BO–June h素, 安定王山王。2015.微软学术服务（MAS）和应用程序的概述。在ww w。
- [39] 和张张。2020.殖民地：语境化语言和知识嵌入。Arxiv预印迹arxi
- [40] J IE tang, jin钊Zhang, I I民Y奥, Juan子L i, Liz航, and Z红SU. 2008. AR net–
- [41] 统计机器翻译梁搜索算法。计算语言学29（2003），97–133。
- [42] Aidan N Gomez, Lukasz Kaiser和Illia Polosukhin。2017.关注是您所需要的。Arxiv预印迹arxiv: 1706.03762（2017）。
- [43] K按W U, J IE tang, and Chen会Zhang. 2018. where have you been? inferring 来自学术社交网络的职业轨迹..在IJCAI。
- [44] 和quoc v le。2019. XLNET：用于语言理解的广泛自动评级预借堡。Arxiv预印迹arxiv: 1906.08237（2019）。
- [45] Peng Zhang Guohao Dai Yu Wang Chang Zhou Hongxia Yang Jie Tang Yukuo Cen, Zhenyu Hou. 2021. CogDL: An Extensive Toolkit for Deep Learning on Graphs. arXiv preprint arXiv:2103.00959 (2021).
- [46] 顾, 严王, 斌邵, 瑞丽, 等。2019.
- [47] J IE Zhang, Y u小dong, Y案Wang, J IE tang, and Ming ding. 2019. prone: fast
- [48] min家Zhang Andy U熊he. 2020. accelerating training of transformer–基于渐进层滴加的基于语言模型。arxiv预印刷品
- [49] Y U套Zhang, fan进Zhang, PEI然Y奥, and J IE tang. 2018. name disambiguation
- [50] Z和ng烟Zhang, X u Han, Z Hi迺I IU, Xinjiang, Mao送sun, and Q UN I IU. 2019. ernie: 与信息实体的语言表示增强。arxiv预印迹arxiv: 1905.07129（2019）。

A EXPERIMENT SUPPLEMENTARY

A.1 Zero-shot Inference

Use of Prompt Word As shown in Table 2, 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 abstract 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 to be 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 $\frac{1}{L^\alpha} \sum_{1 \leq k \leq L} \log P(w_{i_k} | C, w_{i_1}, w_{i_2}, \dots, w_{i_{k-1}})$, where L denotes the length of target entity. When $\alpha = 0$, this equation degrades to the summation version, which is 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 2. 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			
<i>FOS</i>	35.33%, 0.52	32.07%, 0.51	14.85%, 0.36
<i>Venue</i>	18.00%, 0.32	19.30%, 0.33	7.07%, 0.23
<i>Affiliation</i>	12.40%, 0.25	10.83%, 0.23	9.23%, 0.21
OAGBERT			
<i>FOS</i>	49.59%, 0.67	48.08%, 0.66	45.36%, 0.63
<i>Venue</i>	39.00%, 0.57	38.20%, 0.57	36.13%, 0.55
<i>Affiliation</i>	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 have 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 is using 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.

A.2 Supervised Classification

In terms of training, we use the slanted triangular scheduler to adjust the learning rate dynamically and the AdamW optimizer with a maximal learning rate at 2e-5. We run the fine-tuning process for 5 epochs with 10% of the training steps used for warm-up. For each model and each task setting, the averaged accuracy (Hit@1) and standard deviations for 5 runs with different random seeds are reported in Table 4. The number of samples in the classification datasets is shown in Table 11.

A EXPERIMENT SUPPLEMENTARY

A.1 Zero-shot Inference

使用提示单词如表2所示，在FOS推理任务中使用提议的提示单词，结果对斯科格特对解码纸字段（FOS）相当有用。我们猜测它是因为额外的提示单词可以帮助改变预训练模型的重点，同时在蒙面令牌上进行预测。但是，斯科尔特的改善是隶属关系的边缘。在解码场地时，它甚至会伤害性能。这可能是由于迅速单词的选择不当。

对于OAG-BERT，这种技术有限的帮助是我们的预期灰。OAG-BERT而不是使用作为SCIBERT的连续位置，编码实体间位置以区分不同的实体和纸质文本。因此，附加的附加提示单词被视为纸张标题的一部分，并且与OAG-BERT的蒙面实体不相邻。

摘要使用摘要可以大大提高Scibert和OAG-BERT中的模型推理性能。两种模式经常接受预培训过程中的长文本输入，这使得它们自然有利于摘要。此外，摘要可以帮助预训练模型捕获整个纸张的主要思想。

任务比较隶属关系任务似乎比其他两个任务更难。这可能是由于附属名称中包含的弱语义信息。实地研究中的单词可以被视为与纸质内容的共享相同的语言，大多数场地名称也包含诸如“机器学习”或“高能”之类的信息概念词。隶属名称并不总是如此。对于像“Har-Vard大学”或“牛津大学”这样的大学，他们的研究人员可以专注于多个无关域，这很难捕获语言模型。对于公司和研究机构来说，有些人可以专注于一个域，但没有必要以他们的名称具有这样的描述，这也使预培训语言模型混淆。

任务比较隶属关系任务似乎比其他两个任务更难。这可能是由于附属名称中包含的弱语义信息。实地研究中的单词可以被视为与纸质内容的共享相同的语言，大多数场地名称也包含诸如“机器学习”或“高能”之类的信息概念词。隶属名称并不总是如此。对于像“Har-Vard大学”或“牛津大学”这样的大学，他们的研究人员可以专注于多个无关域，这很难捕获语言模型。对于公司和研究机构来说，有些人可以专注于一个域，但没有必要以他们的名称具有这样的描述，这也使预培训语言模型混淆。

讨论等式2中的实体概率，我们使用所有令牌的日志概率之和计算实体日志概率。随着每个令牌的日志概率始终是负的，这种方法似乎对具有更长长度的实体是不公平的。但是，对于基于MLM的模型，编码过程不仅编码 “[掩码]” 令牌，而且还捕获屏蔽实体的长度和每个令牌的位置。因此，如果预先训练的语料库具有比短实体更少的实体更少，则在解码过程中，与短实体中的那些相比，长实体的解码令牌通常会获得更高的概率。

表示目标实体的长度。当 $\alpha = 0$ 时，该方程式化到求和版本，其在以前的任务中使用。当 $\alpha = 1$ 时，此方程式降低到平均版本。

我们使用各种 α 比较不同的平均方法并在零拍摄推理任务上测试它们的性能。我们根据表2选择具有最佳性能的输入功能。对于Scibert，我们使用抽象和提示字进行FOS和Autfiliation推断。我们不使用迅速的词来进行场地推断。对于OAG-BERT，我们只使用摘要随着提示词不起作用。表8中的结果表明，在最多的时间内，使用总结策略显著优于平均策略。简单的平均值（ $\alpha = 1$ ）似乎是最糟糕的选择。然而，对于某些情况，适度平均值（ $\alpha = 0.5$ ）可能是有益的。

表8：在计算实体日志概率的同时使用不同平均方法的结果。据报道，命中率@ 1和MRR。

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

在我们设计的解码程序中进行解码顺序的讨论，我们不会严格遵循古典解码器模型中使用的左右订单。主要原因是对于基于编码器的BERT模型，每个掩码令牌的解码依赖于所有双向上下文信息，而不是仅先前的单词。我们比较表9中使用左右解码和无序解码的性能。

结果表明，对于FOS，没有显著差异在两个解码订单之间，因为候选FOS只有一个或两个令牌。至于地点和隶属关系，事实证明，除了oag-bert使用摘要时，秩序的解码通常比左右解码更好地表现得比左右解码更好。我们还使用表9中的左右解码和提示单词呈现模型的结果，这表明左右解码有时会破坏提示单词的有效性，特别是对于OAG-BERT。

A.2 Supervised Classification

在培训方面，我们使用倾斜三角调度程序动态调整学习速率，以及在2E-5中具有最大学习率的Adamw优化器调整学习速率。我们运行5个时期的微调过程，其中10％的培训步骤用于预热。对于每个模型和每个任务设置，表4中报告了每个模型设置，平均精度（命中@ 1）和5个用不同随机种子运行的标准偏差。表11中示出了分类数据集的样本数。

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

A.3 NLP Tasks

Task Description Among all 15 NLP tasks, 9 tasks concentrate on the field of Biology and Medicine (Bio). Another 4 tasks use paper samples from computer science domain (CS). The rest two tasks involve a mixture of multi-domain data (Multi).

Tasks including NER and PICO require models to make predictions on each token and identify which tokens are part of entities. Some datasets like BC5CDR [26] only need span range identification while other datasets like EBM-NLP [33] also need entity type recognition. For sequence token classification tasks, a Conditional Random Field (CRF) layer is added on top of token outputs from the pre-training model, to better capture the dependencies between sequence labels. The DEP task [20] also uses the token outputs from the pre-training model. The token embeddings, produced by the pre-training model, are fed to a biaffine matrix attention block and used to make further predictions on dependency arc type and direction. The REL and CLS tasks are sequence prediction tasks. The model only needs to make one prediction on the whole sequence. For example, in Paper Field prediction task [38], the model accepts paper title as inputs and output the research fields of that paper. The REL tasks [24, 31], although not directly asking the label of the input sequence, can be reformed into sequence prediction as well. In this type of tasks, the model makes predictions for the entity relation types by categorizing the whole sequence, where the focused

entity pairs are encapsulated with special tokens. The SSC tasks are multi-sequence prediction tasks. Given a list of sentences such as abstract, the model needs to predict the functionality for each internal sentence. These tasks always involve long sequences and also benefits from using CRF layer on top of the sentence embeddings.

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.

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

表9：使用左右解码和订单次序解码顺序的结果。据报道，命中率@ 1和MRR。大于1％的差异的差异均粗略。

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

A.3 NLP Tasks

所有15个NLP任务中的任务说明，9个任务集中在生物学和医学领域（生物）。另外4个任务使用计算机科学域（CS）的纸张样本。其余两个任务涉及多域数据（多）的混合。

包括ner和pico在内的任务需要模型来制作秘密－每个令牌上的速度并确定哪个令牌是实体的一部分。一些数据集如BC5CDR [26]只需要跨度范围识别，而其他数据集如EBM－NLP [33]也需要实体类型识别。对于序列令牌分类任务，在从预训练模型的令牌输出之上添加条件随机字段（CRF）层，以更好地捕获序列标签之间的依赖关系。DEP任务[20]还使用预培训模型的令牌输出。由预训练模型产生的令牌嵌入式被馈送到双基因矩阵注意力块，并用于对依赖弧型和倾向进行进一步的预测。rel和cls任务是序列预测任务。该模型仅需要对整个序列进行一个预测。例如，在纸张字段预测任务[38]中，该模型将纸张标题作为输入和输出该纸张的研究领域。rel任务[24,31]虽然没有直接询问输入序列的标签，但也可以重整为序列预测。在这种类型的任务中，该模型通过对聚焦的整个序列进行分类来使实体再现类型的预测

实体对用特殊令牌封装。SSC任务是多序列预测任务。给定诸如摘要之类的句子列表，模型需要预测每个句子的功能。这些任务始终涉及长序列，并且在句子嵌入的顶部使用CRF层也是有益的。

表10：零拍摄推理任务和监督分类任务中的使用过的候选人的完整列表。

FOS: 艺术，生物学，商业，化学，计算机科学，经济学，工程，环境科学，地理学，地质，历史，MA－Terials科学，数学，医学，哲学，物理学，政治，心理学，社会学
场地: Arxiv: Arxiv: Arxiv: Arxiv分析: Astro－物理, Arxiv: 古典分析和ODES, ARXIV: COMBINATORICS, ARXIV: 计算机视觉和模式识别, ARXIV: 术语: 动态系统, ARXIV: 功能分析, arxiv: arxiv: arxiv: 一般相对论和量子宇宙, arxiv: 地理拓扑, arxiv: arxiv: 高能量物理－实验, arxiv: 高能物理－arxiv: 高能量物理学－理论, arxiv: 学习, arxiv: 材料科学, arxiv: 数学物理学, arxiv: arxiv: arxiv: 核理论, arxiv: 数量理论, arxiv: 数值分析, arxiv: 优化和控制, arxiv: arxiv: arxiv: 量子物理, Arxiv: 代表理论, arxiv: rings和代数, arxiv: 统计力学, arxiv: 强相相关电子
隶属关系: Al Azhar University, Carnegie Mellon University, Carnegie Mellon大学, 中国科学院疾病控制和预防中心, 电力研究院, 复旦大学, 哈佛大学, 哈佛大学, 英特尔, 伊斯兰亚萨德大学, KatholiekeUniversiteit Leuven, Ludwig Maximilian大学慕尼黑大学, Max Planck Society, Mayo Clinic, Moso Soundy University, 北京大学国家科学技术研究委员会, 中国人民大学, 俄罗斯科学院, 西门子, 斯坦福大学, 斯坦福大学, Sun Yat Sen大学, 清华大学, 清华大学, 加州大学伯克利大学牛津大学巴黎大学

表11：监督分类任务中使用的数据集的大小。

任务类别列车验证测试				
FOS	19	152000	19000	19000
Venue	30	24000	3000	3000
Affiliation	30	24000	3000	3000

表12：链路预测中使用的CS异构图的详细信息。

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

Evaluation Metrics We use the same evaluation metrics with the SciBERT [2] paper and the Sequential-Sentence-Classification [5] paper. For NER and PICO tasks, we compare the span-level and token-level macro F1 scores respectively, except using micro-F1 for ChemProt [24]. For REL, CLS, and SSC tasks, we compare sentence-level macro F1 scores. For the DEP task, we compare LAS (labeled attachment score) and UAS (unlabeled attachment score).

Hype-parameters In SciBERT, the authors claimed that the best results for most downstream tasks were produced by fine-tuning 2 or 4 epochs and using 2e-5 learning rate after searching between 1 to 4 epochs with a maximum learning of 1e-5, 2e-5, 3e-5, 5e-5, as stated in [30]. In our experiments, we follow the same settings and select the optimal hyper-parameters on validation sets and report the corresponding test sets results.

Table 13: The performance of vanilla OAG-BERT with and without training on 512-token samples. All results in this table were produced by fine-tuning with 2 epochs and 2e-5 learning rates.

Task	Dataset	Vanilla OAG-BERT		Gain
		w/o 512	w/ 512	
NER	BC5CDR	89.62 ^{.16}	89.33 ^{.12}	-0.29
	NCBI-disease	87.63 ^{.62}	87.92 ^{1.08}	+0.29
	SciERC	67.64 ^{.52}	67.19 ^{.34}	-0.45
REL	ChemProt	77.50 ^{1.99}	77.99 ^{2.50}	+0.49
	SciERC	69.87 ^{1.51}	69.88 ^{.77}	+0.01
SSC	NICTA-piboso	77.62 ^{.87}	80.01 ^{.24}	+2.39
	CSAbstruct	72.65 ^{.40}	82.30 ^{.47}	+9.65

Pre-training on 512-token samples During fine-tuning on NLP tasks, we also observe that the pre-training on inputs with 512 tokens is essential for the SSC tasks with up to 10% performance boost, which is much larger than the performance boost for other types of tasks as shown in Table 13. It is because the SSC tasks require the model to comprehend multiple sentences in a long paragraph rather than a single sentence in other tasks.

评估指标我们使用与SCIBERT [2]纸张的相同的评估指标和顺序句子分类[5]纸。对于Ner和Pico任务，我们可以分别比较Span级和令牌级宏F1分数，但是除了ChemProt的Micro-F1 [24]。对于Rel，Cls和SSC任务，我们比较句子级宏F1分数。对于DEP任务，我们比较LAS（标记为附件得分）和UAS（未标记的附件得分）。

作者声称在SCIBERT中的炒作参数声称，大多数下游任务的最佳效果是通过微调2或4个时期和使用2E-5学习率来搜索1到4个时期的最大学习1E-5后产生的，如[30]所述，2E-5,3E-5, 5E-5。在我们的实验中，我们按照相同的设置，然后在验证集上选择最佳超参数，并报告相应的测试集结果。

表13：Vanilla OAG-BERT的性能与512-令牌样品上的训练。该表中的所有结果都是通过用2个时期和2E-5学习率进行微调来生产的。

Task	Dataset	Vanilla OAG-BERT		Gain
		w/o 512	w/ 512	
NER	BC5CDR	89.62 ^{.16}	89.33 ^{.12}	-0.29
	NCBI-disease	87.63 ^{.62}	87.92 ^{1.08}	+0.29
	SciERC	67.64 ^{.52}	67.19 ^{.34}	-0.45
REL	ChemProt	77.50 ^{1.99}	77.99 ^{2.50}	+0.49
	SciERC	69.87 ^{1.51}	69.88 ^{.77}	+0.01
SSC	NICTA-piboso	77.62 ^{.87}	80.01 ^{.24}	+2.39
	CSAbstruct	72.65 ^{.40}	82.30 ^{.47}	+9.65

在对NLP任务的微调期间进行512-令牌样本的预先训练，我们还观察到使用512令牌的输入进行预训练对于最高可达10%性能提升的SSC任务至关重要，这远远大于性能如表13所示的其他类型任务提升。因此，由于SSC任务要求模型在长段中而不是其他任务中的单个句子来理解多个句子。