通过图卷积网络进行跨语言知识图对齐

摘要

DBpedia和YAGO等多语言知识图谱（KG）包含多种不同语言的实体的结构化知识，它们是跨语言AI和NLP应用的有用资源。跨语言知识图谱对齐是将实体与不同语言的对应实体进行匹配的任务，是丰富多语言知识图谱中跨语言链接的重要方式。在本文中，我们提出了一种通过图卷积网络（GCN）进行跨语言知识图谱对齐的新方法。给定一组预先对齐的实体，我们的方法训练GCN将每种语言的实体嵌入到统一的向量空间中。实体对齐是根据嵌入空间中实体之间的距离发现的。嵌入可以从实体的结构和属性信息中学习，并将结构嵌入和属性嵌入的结果结合起来以获得准确的对齐。在对齐真实多语言知识图谱的实验中，与其他基于嵌入的知识图谱对齐方法相比，我们的方法获得了最佳性能。

# 1Introduction

知识图谱（KG）以机器可读的格式表示人类知识，正在成为人工智能和自然语言处理领域许多应用的重要基础。如果要跨语言应用程序，诸如DBpedia(Bizeretal.,2009)、YAGO(Suchaneketal.,2008;Rebeleetal.,2016)和BabelNet(NavigliandPonzetto,2012)等多语言知识图谱尤其有价值被建造。除了用每种不同语言编码的知识之外，多语言知识图谱还包含丰富的跨语言链接，与不同语言中的等效实体相匹配。跨语言链接对于弥合多语言知识库中的语言差距发挥着重要作用；然而，在大多数多语言知识图谱中，并非所有等效实体都通过跨语言链接进行连接。因此，越来越多的研究工作研究跨语言知识图谱对齐问题，旨在自动匹配多语言知识图谱中不同语言的实体。

传统的跨语言知识图谱对齐方法要么依赖机器翻译技术，要么定义各种与语言无关的特征来发现跨语言链接。最近，人们提出了几种基于嵌入的方法用于跨语言KG对齐，包括MTransE（Chen等人，2017）和JAPE（Sun等人，2017）。给定两个知识图谱和一组预先对齐的实体，基于嵌入的方法将实体投影到低维向量空间中；实体根据其向量表示的计算进行匹配。遵循与上述非常相似的想法，JE（Haoetal.,2016）和ITransE（Zhuetal.,2017）是基于嵌入的方法，用于在异构知识图谱之间匹配实体，它们也可以解决跨语言问题公斤对齐。上述基于嵌入的方法无需机器翻译或特征工程即可实现良好的性能。

然而，我们发现上述方法都试图在一个统一的优化问题中对跨语言知识和单语言知识进行联合建模。在优化过程中必须仔细平衡两种知识的损失。例如，JE、MTransE和ITransE都使用超参数来加权其模型损失函数中实体对齐的损失；JAPE使用预先对齐的实体将两个KG合二为一，并在其损失函数中为负样本的分数添加权重。在上述方法中，实体的嵌入必须对知识图谱中的结构信息和实体的等价关系进行编码。此外，现有模型尚未充分利用实体的属性（例如，人的年龄、国家的人口）。MTransE和ITransE不能使用KG中的归属信息；尽管JAPE在模型中包含了属性类型，但实体的属性值被忽略。我们相信考虑属性值可以进一步改善KG对齐的结果。

有了上述观察，我们提出了一种新的基于嵌入的KG对齐方法，该方法使用图卷积网络（GCN）直接建模实体之间的等效关系。GCN是一种直接对图结构数据进行操作的卷积网络；它通过编码有关节点邻域的信息来生成节点级嵌入。KG中两个等价实体的邻接通常包含其他等价实体，因此我们选择GCN来生成实体的邻域感知嵌入，用于发现实体对齐。我们的方法还可以提供一种简单有效的方法来将实体的属性值包含在对齐模型中。更具体地说，我们的方法具有以下优点：

我们的方法使用每个KG中的实体关系来构建GCN的网络结构，并且在模型训练中仅考虑实体之间的等效关系。我们的方法模型复杂度较小，并且可以实现令人鼓舞的对齐结果。

我们的方法只需要预先对齐的实体作为训练数据，并且不需要知识图谱之间的任何预先对齐的关系或属性。

我们的方法有效地结合了实体关系和实体属性来改善对齐结果。

在对齐真实多语言知识图谱的实验中，与基线方法相比，我们的方法获得了最佳性能。

本文的其余部分组织如下，第2节回顾了一些相关工作，第3节介绍了一些背景知识，第4节描述了我们提出的方法，第5节介绍了评估结果，第6节是结论和未来的工作。

# 2RelatedWork

## 2.1KG Embedding

在过去的几年里，人们在KG嵌入问题上做了很多工作。知识图谱嵌入模型将知识图谱中的实体和关系嵌入到低维向量空间中，同时保留原始知识。嵌入通常是通过最小化知识图谱中所有实体和关系的全局损失函数来学习的，它可以进一步用于关系预测、信息提取和其他一些任务。TransE是一种代表性的知识图谱嵌入方法（Bordesetal.,2013），它将实体和关系投影到同一个向量空间中；如果三元组(h,r,t)成立，TransE希望h+r≈t。通过最小化训练集上基于边距的排名标准来学习嵌入。TransE模型简单但功能强大，在链路预测和三元分类问题上取得了可喜的结果。为了进一步改进TransE，人们提出了几种基于TransE的增强模型，包括TransR(Linetal.,2015)、TransH(Wangetal.,2014)和TransD(Jietal.,2015)等。关系翻译的表示，后来的方法以增加模型复杂性为代价获得了更好的性能。还有许多其他的KG嵌入方法，最近的调查（Wangetal.,2017;Nickeletal.,2016）给出了详细的介绍和比较。

## 2.2Embedding-based KG Alignment

在这里，我们介绍与我们最相关的KG对齐方法，并讨论我们的方法与它们之间的主要区别。

JE（Haoetal.,2016）联合学习多个知识图谱在统一向量空间中的嵌入，以对齐知识图谱中的实体。JE使用一组种子实体对齐来连接两个KG，然后使用修改后的TransE模型来学习嵌入，该模型在其全局损失函数中添加了实体对齐的损失。

MTransE(Chenetal.,2017)使用TransE在单独的嵌入空间中对每个KG的实体和关系进行编码；它还为每个嵌入向量提供到其他空间中的跨语言对应向量的转换。MTransE的损失函数是两个组件模型（即知识模型和对齐模型）损失的加权和。为了训练对齐模型，MTransE需要一组两个KG的对齐三元组。

JAPE（Sunetal.,2017）结合了结构嵌入和属性嵌入来匹配不同知识图谱中的实体。结构嵌入遵循TransE模型，该模型学习两个KG的叠加图中实体的向量表示。属性嵌入遵循Skip-gram模型，旨在捕获属性的相关性。为了得到理想的结果，JAPE需要提前对齐两个KG的关系和属性。

ITransE（Zhuetal.,2017）是一种针对多个知识图谱的联合知识嵌入方法，也适用于跨语言知识图谱对齐问题。ITransE首先学习TransE之后的实体和关系嵌入；然后它学习根据一组种子实体对齐将不同知识图谱的知识嵌入映射到联合空间中。ITransE通过使用新发现的实体对齐来更新实体的联合嵌入来执行迭代实体对齐。ITransE要求KG之间共享所有关系。

上述方法遵循类似的框架来匹配不同知识图谱中的实体。它们都依赖TransE模型来学习实体嵌入，然后定义对齐实体嵌入之间的某种转换。与这些方法相比，我们的方法使用了完全不同的框架；它使用GCN将实体嵌入到统一的向量空间中，其中对齐的实体预计尽可能接近。我们的方法仅关注两个知识图谱中的实体匹配，并且不学习关系的嵌入。MTransE、JAPE和ITransE都要求在KG中对齐或共享关系；我们的方法不需要这种先验知识。

# 3Problem Formulation

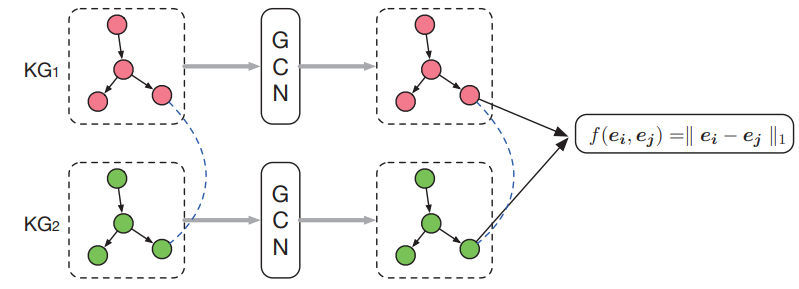
KG将现实世界实体的知识表示为三元组。这里我们考虑知识图谱中的两种三元组：关系三元组和归因三元组。关系三元组表示实体之间的关系，其形式为hentity1、relation、entity2i。属性三元组描述实体的属性，其形式为hentity、attribute、valuei。例如在YAGO的数据中，gradedFrom是一个关系，(AlbertEinstein,graduateFrom,ETHZurich)是一个关系三元组；deadOnDate是一个属性，(AlbertEinstein,deadOnDate,1955)是一个归因三元组。关系三元组和归因三元组都描述了有关实体的重要信息，我们将在跨语言KG对齐的任务中考虑它们。

形式上，我们将KG表示为G=(E,R,A,TR,TA)，其中E、R、A分别是实体、关系和属性的集合；TR⊂E×R×E是关系三元组的集合，TA⊂E×A×V是属性三元组的集合，其中V是属性值的集合。

设G1=(E1,R1,A1,TR1,TA1)和G2=(E2,R2,A2,TR2,TA2)为不同语言的两个KG，为G1之间预先对齐的实体对的集合和G2。我们将跨语言知识图谱对齐的任务定义为基于现有的实体对齐找到新的实体对齐。在DBpedia和YAGO等多语言知识图谱中，其中的跨语言链接可用于构建预对齐实体对的集合。已知的实体比对在KG比对过程中用作种子或训练数据。

# 4The Proposed Approach

我们提出的方法的框架如图1所示。给定不同语言的两个知识图谱G1和G2，以及它们之间的一组已知对齐实体对，我们的方法根据基于GCN的实体嵌入自动找到新的实体对齐。我们方法的基本思想是使用GCN将不同语言的实体嵌入到统一的向量空间中，其中等效实体预计尽可能接近。通过将预定义的距离函数应用于实体的GCN表示来预测实体对齐。



**图1：我们的方法框架（蓝色虚线连接两个KG中的等效实体）**

## 4.1GCN-based Entity Embedding

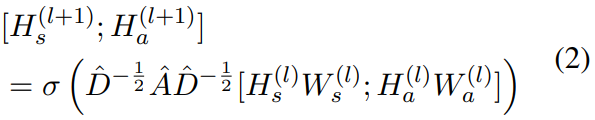
GCN（Bruna等人，2014；Henaff等人，2015；Defferrard等人，2016；Kipf和Welling，2017）是一种直接对图数据进行操作的神经网络。GCN允许对预测管道进行端到端学习，其输入是任意大小和形状的图。GCN的输入是节点的特征向量和图的结构；GCN的目标是学习输入图上的特征函数并产生节点级输出。GCN可以将节点邻域的信息编码为实值向量，通常用于分类或回归。在解决KG对齐问题时，我们假设（1）等价实体往往具有相似的属性，（2）等价实体通常与其他一些等价实体相邻。GCN可以将属性信息和结构信息结合在一起，因此我们的方法使用GCN将实体投影到低维向量空间中，其中等效实体彼此靠近。

GCN模型由多个堆叠的GCN层组成。GCN模型第l层的输入是顶点特征矩阵，其中n是顶点数量，d(l)是第l层中的特征数量。第l层的输出是通过以下卷积计算得到的新特征矩阵H(l+1)：



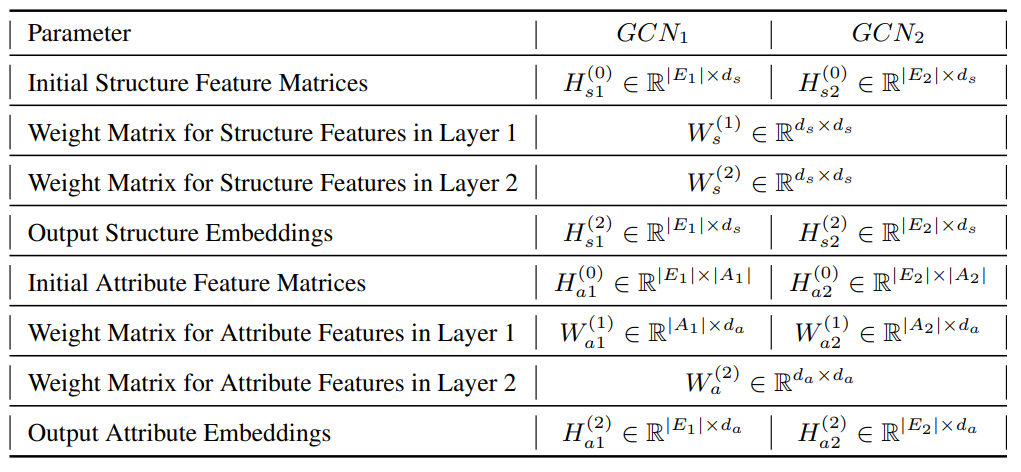
这里σ是激活函数；A是一个n×n的连通矩阵，表示图的结构信息；A^=A+I，I为单位矩阵；Dˆ为Aˆ的对角节点度矩阵；是GCN中第l层的权重矩阵，d(l+1)是新顶点特征的维数。

结构和属性嵌入。在我们的方法中，GCN用于将两个KG的实体嵌入到统一的向量空间中。为了利用实体的结构和属性信息，我们的方法为GCN层中的每个实体分配两个特征向量：结构特征向量hs和属性特征向量ha。在输入层，在训练过程中随机初始化和更新；是实体的属性向量，在模型训练过程中是固定的。设Hs和Ha为所有实体的结构和属性特征矩阵，我们将卷积计算重新定义为：



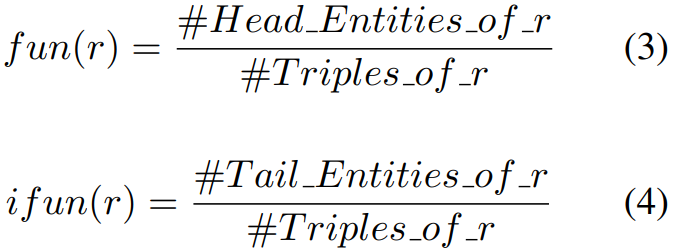
其中和分别是第l层结构特征和属性特征的权重矩阵；[;]表示两个矩阵的串联。激活函数σ选择为ReLU(·)=max(0,·)。

**表1：两个GCN的参数**

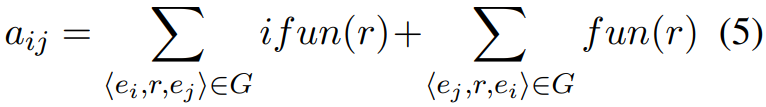


型号配置。更具体地说，我们的方法使用两个2层GCN，每个GCN处理一个KG来生成其实体的嵌入。根据第3节的定义，我们将两个KG表示为G1=(E1,R1,A1,TR1,TA1)和G2=(E2,R2,A2,TR2,TA2)；并将它们对应的GCN模型记为GCN1和GCN2。对于实体的结构特征向量，我们在GCN1和GCN2的所有层中将特征向量的维数设置为ds；两个GCN模型共享两层结构特征的权重矩阵和。对于实体的属性向量，我们将输出特征向量的维度设置为da。由于两个KG可能具有不同数量的属性（即|A1|≠|A2|），因此两个GCN模型中输入属性特征向量的维度是不同的。每个GCN模型的第一层将输入的属性特征向量变换为大小为da的向量；两个GCN模型生成相同维度的属性嵌入。表1概述了我们方法中两个GCN的参数。两个GCN的最终输出是实体的(ds+da)维嵌入，进一步用于发现实体对齐。

连接矩阵的计算。在GCN模型中，连接矩阵A定义了卷积计算中实体的邻域。对于无向图，可以直接使用邻接矩阵As。但KG是关系多重图，实体通过类型关系连接。因此，我们设计了一种特殊的方法来计算KG的A；我们让aij∈A表示比对信息从第i个实体传播到第j个实体的程度。考虑到两个实体通过不同的关系连接到对齐的实体（例如，hasParent与hasFriend），两个实体等效的概率有很大差异。因此，我们为每个关系计算两个度量，称为函数和逆函数：

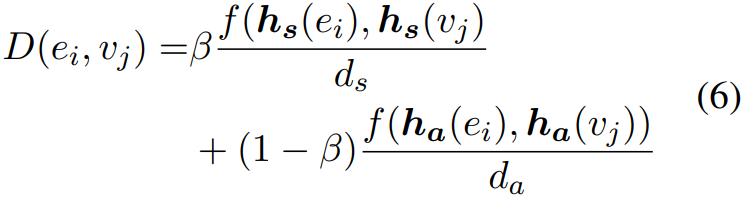


其中#Triples\_of\_r是关系r的三元组数量；#Head\_Entities\_of\_r和#Tail\_Entities\_of\_r分别是r的头实体和尾实体的数量。为了衡量第i个实体对第j个实体的影响，我们将aij∈A设置为：



## 4.2Alignment prediction

实体对齐是根据GCN表示空间中两个KG的实体之间的距离来预测的。对于G1中的实体ei和G2中的实体vj，我们计算它们之间的以下距离度量：

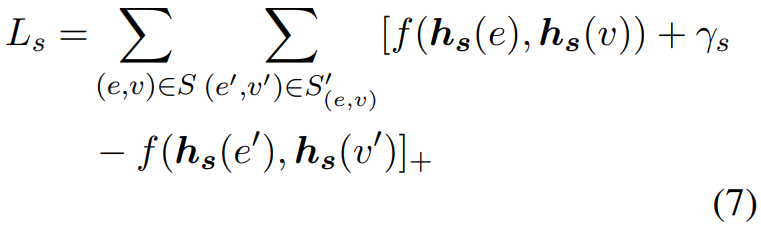


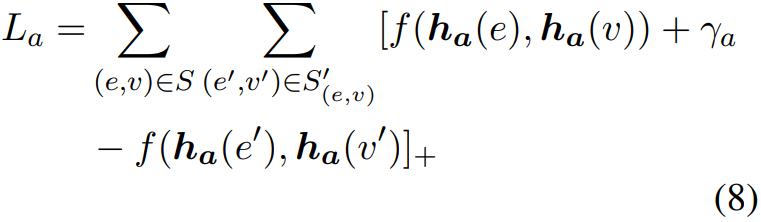
其中、hs(·)和ha(·)分别表示实体的结构嵌入和属性嵌入；ds和da是结构嵌入和属性嵌入的维度；β是一个超参数，用于平衡两种嵌入的重要性。

对于等效实体，该距离预计较小，对于非等效实体，该距离预计较大。对于G1中的特定实体ei，我们的方法计算ei与G2中所有实体之间的距离，并返回排名实体的列表作为候选对齐。也可以从G2到G1进行对齐。在实验中，我们报告了KG对齐两个方向的结果。

## 4.3Model Training

为了使GCN能够在向量空间中尽可能接近地嵌入等效实体，我们使用一组已知的实体对齐S作为训练数据来训练GCN模型。模型训练是通过最小化以下基于边际的排名损失函数来执行的：





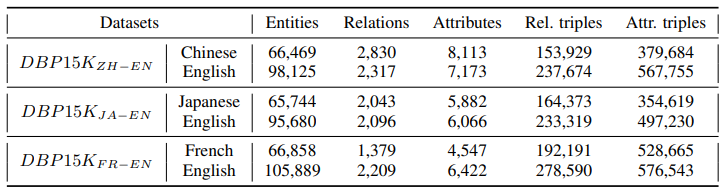
其中[x]+=max{0,x}，表示通过破坏(e,v)构造的负实体对齐集合，即将e或v替换为G1或G2中随机选择的实体；γs,γa>0是分隔正负实体对齐的边际超参数。Ls和La分别是结构嵌入和属性嵌入的损失函数；它们彼此独立，因此单独优化。我们采用随机梯度下降（SGD）来最小化上述损失函数。

# 5Experiment

## 5.1Datasets

我们在实验中使用DBP15K数据集，该数据集由Sun等人构建。（2017）。数据集是从DBpedia生成的，DBpedia是一个大型多语言知识库，包含不同语言版本之间丰富的语言间链接。DBpedia的中文、英文、日文和法文版本的子集是按照一定的规则选择的。表2概述了数据集的详细信息。每个数据集包含两个不同语言的知识图谱的数据以及连接两个知识图谱中等效实体的15000个跨语言链接。在实验中，使用已知的等效实体对进行模型训练和测试。

**表2：数据集详细信息**



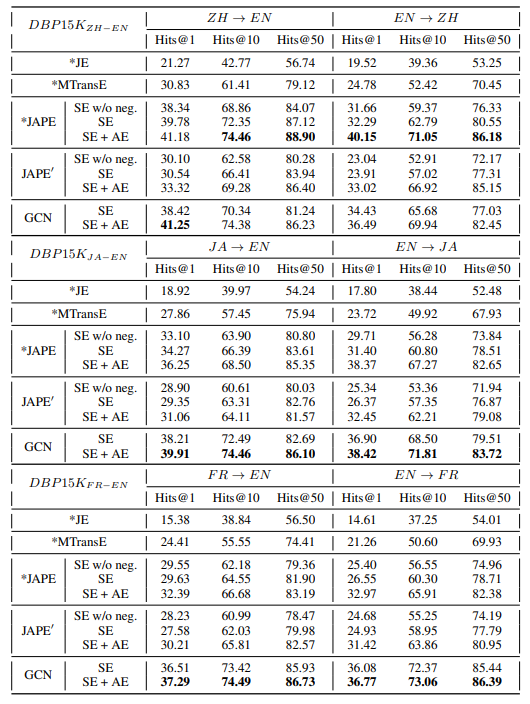
## 5.2Experiment Settings

在实验中，我们将我们的方法与JE、MTransE和JAPE进行了比较。我们还构建了JAPE‘，这是JAPE的一个变体，它不使用预先对齐的关系和属性。由于ITransE方法执行迭代对齐，并且需要两个知识图谱具有相同的关系，因此我们不将其包含在比较中。每个数据集中的中介语链接被用作实体对齐的黄金标准。对于所有比较的方法，我们使用30%的跨语言链接进行训练，70%用于测试；所有方法的训练和测试的划分都是相同的。我们使用Hits@k作为评估指标来评估所有方法的性能。Hits@k衡量排名前k个候选者中正确对齐的实体的比例。对于我们方法的参数，我们设置ds=1000，da=100；损失函数中的边际γs=γa=3，距离度量中的β根据经验设置为0.9。

## 5.3Results

表3显示了DBP15K数据集上所有比较方法的结果。我们报告每个数据集上方法的Hits@1、Hits@10和Hits@50。因为我们使用与(Sunetal.,2017)相同的数据集，所以JE、MTransE和JAPE的结果来自(Sunetal.,2017)。对于JAPE和JAPE0，它们各自具有三种变体：无负三元组的结构嵌入（SEw/oneg.）、结构嵌入（SE）、结构和属性联合嵌入（SE+AE）。我们使用GCN(SE)和GCN(SE+AE)来表示我们方法的两种变体：一种仅使用关系三元组来执行结构嵌入，另一种使用关系三元组和属性三元组来执行结构和属性嵌入。

**表3：跨语言KG比对结果比较（\*标记从（Sunetal.,2017）获得的结果）**



**GCN(SE)vs.GCN(SE+AE)**

我们首先比较GCN(SE)和GCN(SE+AE)的结果，看看归因信息对KG对齐任务是否有帮助。根据结果，在我们的方法中添加属性确实会带来稍微更好的结果。改进范围从1%到10%，这与JAPE(SE)相对于JAPE(SE+AE)的改进非常相似。这表明知识图谱对齐主要依靠知识图谱中的结构信息，但属性信息仍然有用。我们的方法使用相同的框架来嵌入结构和属性信息，两种嵌入的组合可以有效地工作。

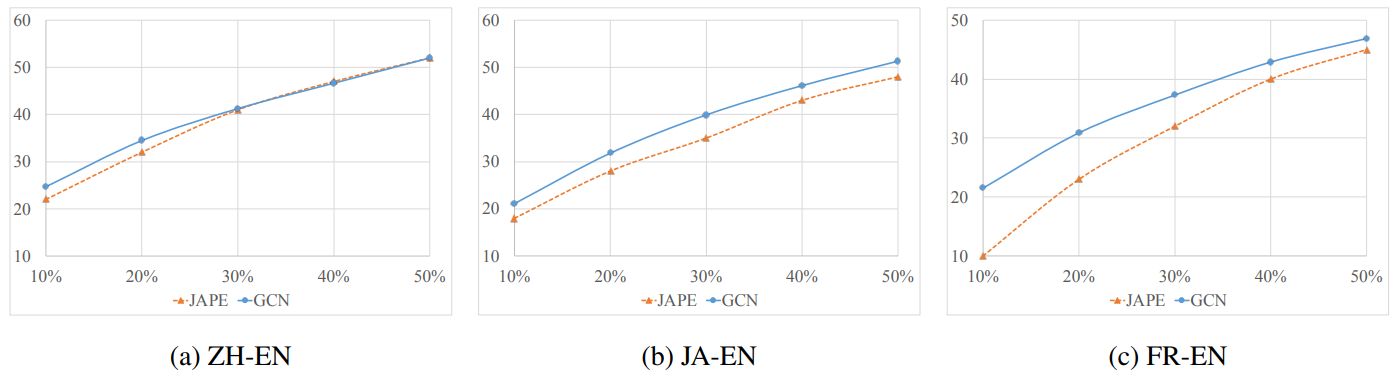
**GCN(SE+AE)vs.Baselines**

在DBP15KZH−EN数据集上，JAPE(SE+AE)表现最好，获得五个最佳Hits@k值；我们的方法GCN(SE+AE)在ZH→EN的对齐方向上获得最佳Hits@1。GCN(SE+AE)和JAPE的结果在ZH→EN方向上关于Hits@1和Hits@10的结果非常接近。在EN→ZH的排列方向上，JAPE(SE+AE)比GCN(SE+AE)高出约2-3%。但应该注意的是，JAPE使用额外的对齐关系和属性作为输入，而我们的方法不使用这些先验知识。如果与JAPE'相比，GCN(SE+AE)的表现优于重新评级Hits@1和Hits@10。与JE和MTransE相比，GCN(SE+AE)在大多数情况下优于它们10%以上。在DBP15KJA−EN和DBP15KFR−EN的数据集上，GCN(SE+AE)在所有Hits@k度量方面优于所有比较方法。即使与使用额外关系和属性对齐的JAPE相比，GCN(SE+AE)仍然获得比它更好的结果。

与所有基线相比，GCN(SE)和GCN(SE+AE)均显着优于JE和MTransE。在所有基线中，JAPE是最强的；这可能是因为它能够同时使用关系三元组和归因三元组，以及它所消耗的关系和属性的额外对齐。我们的方法在两个数据集上取得了比JAPE更好的结果；尽管JAPE的表现比我们的方法更好，但它们的结果之间的差异很小。如果两个知识图谱之间不存在现有的关系和属性对齐，我们的方法将比JAPE具有明显的优势。

**GCNvs.JAPEusingdifferentsizesoftrainingdata**

为了研究训练集的大小如何影响我们方法的结果，我们通过使用不同数量的预对齐实体作为训练数据，进一步将我们的方法与JAPE进行比较。对于JAPE，预对齐的实体用作种子，使其向量重叠。在我们的方法中，所有预对齐的实体都用于训练GCN模型。直观上，使用的预对齐实体越多，GCN和JAPE应该获得更好的结果。



**图2：GCN和JAPE使用不同大小的训练数据（横坐标：训练数据中使用的预对齐实体的比例；纵坐标：Hits@1）**

这里我们使用不同比例的预对齐实体作为训练数据，范围为10%到50%，步长为10%；所有其余的预对齐实体都用于测试。图2显示了三个数据集中两种方法的Hits@1。它表明，随着训练数据大小的增加，两种方法都表现得更好。除了使用40%预对齐实体作为图2(a)中的训练数据外，我们的方法始终优于JAPE。特别是在日语与英语、法语与英语的对齐任务中，我们的方法比JAPE具有明显的优势。

# 6ConclusionandFutureWork

本文提出了一种新的基于嵌入的知识图谱对齐方法，该方法根据GCN学习的实体嵌入来发现实体对齐。我们的方法可以利用知识图谱中的关系三元组和归因三元组来发现实体对齐。我们在真实多语言知识图谱的数据上评估我们的方法，结果表明我们的方法相对于比较基线的优势。

在未来的工作中，我们将探索更先进的用于KG对齐任务的GCN模型，例如关系GCN（Schlichtkrull等，2017）和图注意力网络（GAT）（Velickovic等，2017）。此外，如何在我们的方法框架中迭代地发现新的实体对齐是我们未来研究的另一个有趣的方向。