一种基于实体描述和知识向量相似度的跨语言实体对齐模型

康世泽,吉立新,刘树新,丁悦航
(战略支援部队信息工程大学,河南郑州 450002)

摘 要: 跨语言实体对齐旨在找到不同语言知识图谱中指向现实世界同一事物的实体.传统的跨语言实体对齐方法通常仅依靠知识图谱内部的结构信息,但实际上一些知识图谱提供的实体描述信息也可以被利用.本文提出了一种结合知识图谱的内部结构和实体描述信息共同进行跨语言实体对齐的模型.该模型首先通过训练基于知识图谱结构信息的知识向量找到可能被对齐的实体对,再结合实体描述信息利用改进后的共享参数模型选出最终的对齐实体,最后通过迭代对齐的方法重复前两个步骤找到更多的对齐实体直到训练结束.实验结果表明,与基准算法相比,本文所提模型在跨语言实体对齐任务上可以取得相对不错的结果.

关键词: 跨语言实体对齐;知识向量;跨语言实体描述相似度
中图分类号: TP393.1 **文献标识码:** A **文章编号:** 0372-2112 (2019)09-1841-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.09.004

Cross-Lingual Entity Alignment Model Based on the Similarities of Entity Descriptions and Knowledge Embeddings

KANG Shi-ze, JI Li-xin, LIU Shu-xin, DING Yue-hang
(*Information Engineering University, Zhengzhou, Henan* 450002, *China*)

Abstract: Cross-lingual entity alignment aims to find entities in knowledge graphs of different languages that point to the same objects in the real world. Traditional cross-lingual entity alignment methods usually rely solely on the internal structure information of the knowledge graph, but in fact entity description information provided by some knowledge graphs can also be utilized. This paper proposes an entity alignment model that combines the internal structure information of the knowledge graph with the entity description information for cross-lingual entity alignment. The model first finds the entity pairs that may be aligned by training the knowledge embeddings based on the structure information of the knowledge graph, and then uses entity descriptions to select the final aligned entity pairs based on the improved optimal alignment similarity model. Finally, the model iteratively align the first two steps to find more aligned entity pairs until the end of the training. The experimental results show that compared with the benchmark algorithms, the proposed model can achieve relatively good results in cross-lingual entity alignment task.

Key words: cross-lingual entity alignment; knowledge embeddings; cross-lingual description similarity

1 引言

随着信息的爆炸性增长,人们提出知识图谱来结构化地组织知识.知识图谱基本的存储单元为三元组(头实体,关系,尾实体)[1].头实体和尾实体代表现实世界的各种概念,关系用来连接实体并刻画其间的关联.

多语言知识图谱有 Dbpedia[2]、WordNet[3] 和 Yago[4] 等.这些知识图谱对于知识的全球化共享具有重要意义,但仍存在一些问题:(1)不同语言的知识图谱具有不同的知识分布,非英语知识图谱的信息量通常都比英语知识图谱的少且稀疏.(2)尽管目前有些多语言知识图谱具有匹配不同语言版本相同实体的跨语言链接(ILLs, Inter-Lingual Links),但这些链接通常只能覆

收稿日期:2018-10-11;修回日期:2019-04-09;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61521003, No.61601513)

盖少于15%的实体.

基于上面的问题,有必要将不同语言知识图谱的相同实体对齐,从而使这些知识图谱的信息互补并投入到一些智能应用[5,6]之中. 传统的跨语言实体对齐方法通常基于机器翻译技术[7]或多语言词汇网络[8],这些方法的精度通常严重依赖翻译质量或词汇网络的覆盖率.

最近基于向量的方法也被使用在了知识图谱的表示中,利用该方法获得的向量蕴含知识图谱中不同实体间的语义(semantic)关系. 基于向量的方法可以利用多语言图谱中的ILLs来连接不同语言版本的知识图谱并获取更多的对齐实体. 目前已有一些基于向量方法的跨语言实体对齐模型,它们大部分都只利用了知识图谱的结构信息,但还有一些其它的信息例如多语言知识图谱中存在的描述信息[6]可以利用. 然而衡量不同语言实体描述信息的相似度比较困难,因为:(1)衡量不同语言句子的相似度本身就是难点;(2)不同语言的实体描述信息都是独立编写的,所蕴含的信息不完全相同,从而进一步加大了衡量不同语言实体描述信息相似度的困难.

为了解决以上所提挑战,本文提出同时度量实体结构相似度和实体描述信息相似度的办法来进行跨语言实体对齐. 本文提出的模型首先基于共享参数模型利用ILLs进行知识向量的训练,当训练结果稳定后在知识向量相似度较大的实体间使用基于跨语言词向量的最优对齐模型判断实体描述信息的相似度并选出一些新对齐实体,之后再重新初始化知识向量并基于ILLs和新对齐实体重新训练知识向量,重复这样的过程直到模型达到停止训练的条件. 如下图1所示为本文所提模型在一个训练周期里的示意图.

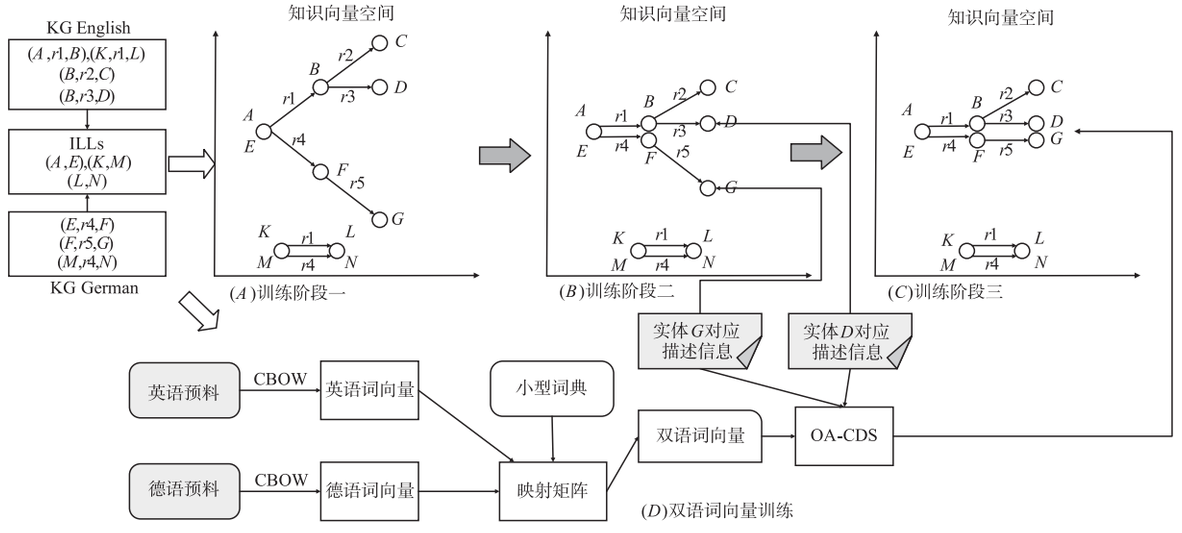

图1 给定一个英语知识图谱、一个德语知识图谱及它们之间的ILLs, B和F的知识向量在完成第一个阶段的训练后相等。之后使用最优对齐模型计算D和G对应实体描述信息的相似度来决定D和G是否能够对齐

2 相关工作

2.1 知识图谱的向量化表示

近几年知识图谱的向量表示取得了很多进展. 表示学习的代表模型包括距离模型[9]、单层神经网络模型[10]、和翻译模型等. 目前主流的模型是翻译模型TransE[11],在其之后的工作大多都是针对TransE所进行的改进. 例如针对TransE无法处理复杂关系的问题,文献[12]提出了TransH模型;针对知识图谱中实体和关系存在的异质性和不平衡性的问题,文献[13]中提出了TranSparse模型.

知识图谱的表示学习除了三元组以外,还可以融合实体对应的描述信息、图片信息以及特定知识图谱的逻辑规则[14]等信息增强表示学习的性能.

2.2 跨语言实体对齐

目前已有一些基于向量表示的跨语言实体对齐模型,最具代表性的是文献[15]中提出的MTransE. 该模型基于TransE学习不同语言知识向量空间之间的转换. 文献[16]在MTransE的基础上提出了ITransE模型,该模型提供了参数共享和迭代对齐的方法来提高对齐的性能. 文献[17]提供了一种基于bootstrapping的方法来减少迭代对齐过程中的错误. 以上这几种方法都没有融入其它信息辅助实体对齐. JAPE模型[18]在学习不同语言知识图谱的知识向量时还考虑了实体的属性信息. 与以上的方法不同,本文在考虑多语言知识图谱的结构信息的同时还融入了实体的描述信息.

3 问题定义

在一个知识图谱(Knowledge Graph,KG)中,每条知

识可以表示为三元组(h,r,t),其中 h 和 t 分别表示头实体和尾实体,实体的集合可以表示为 E;r 表示实体之间的关系. h,r,t 所对应的向量表示分别为 $\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}$. I 表示语言集合,I_2表示任意两种语言之间的无序组合. 任意 $i\in I$ 所对应的 KG 可以表示为 KG_i. 对于一对语言组合$(i_1,i_2)\in I^2$,$\mathrm{ILL}(i_1,i_2)$表示KG_{i_1}和KG_{i_2}之间的已对齐实体即跨语言链接 ILLs,例如(Culture,Kultur)为 ILL(English,German)中的一组已对齐实体对,其中 Culture $\in \mathrm{KG}_{\mathrm{English}}$,Kulture $\in \mathrm{KG}_{\mathrm{German}}$. 大部分实体 $e_j\in E_i$ 都存在一段文字描述 d_j,这些描述的集合为 D_i. 对于任一 $i\in I$,其 D_i 都存在一个词汇表 V_i.

4　方法

本文在这一部分首先介绍了如何计算跨语言描述信息的相似度,之后介绍了如何训练跨语言知识向量,最后提出了迭代对齐的方法提升跨语言实体对齐的性能.

4.1　跨语言描述相似度模型

本文提出了一种非监督的跨语言文本相似度度量方法. 如图 1(D)部分所示,该方法首先根据各语言的语料独立训练各自的词向量,再学习不同语言词向量之间的映射使源语言的词向量空间在转换后与目标语言的词向量空间尽量接近,最后根据训练好的跨语言词向量使用改进后的最优对齐方法来度量不同语言实体描述之间的相似度.

4.1.1　跨语言词向量

本文首先使用 CBOW(Continuous Bag Of Words)模型训练各语言的词向量,再采用线性映射的方法训练源语言和目标语言词向量间的线性映射. 给定一组小型的词典,其对应的词向量表示为$\{\boldsymbol{x}_i,\boldsymbol{z}_i\}_{i=1}^{n}$($x_i$为源语言词向量,$z_i$为目标语言词向量),则源语言和目标语言间的线性映射 $\boldsymbol{W}$ 为:

$$\min_{W}\sum_{i=1}^{n}\ \|\boldsymbol{z}_i-\boldsymbol{W}\boldsymbol{x}_i\|^2 \tag{1}$$

完成训练之后,源语言词向量空间的任一向量 $\boldsymbol{x}$ 可以通过式 $\boldsymbol{z}^*=\boldsymbol{W}\boldsymbol{x}$ 投射到目标语言的空间. 源语言和目标语言词向量的相似度可以采用余弦测度进行计算.

4.1.2　最优对齐相似度模型(OA-CDS,Optimal Alignment Cross-lingual Description Similarity Model)

本文基于最优对齐模型[19]结合实体描述信息中的有效特征提出了改进的最优对齐模型. 最优对齐模型是一种仅基于跨语言词向量的鲁棒性方法,其目标是找到源语言文本和目标语言文本词语之间的最优对齐,具体的办法是使对齐的词语对的相似度值之和达到最大. 本文在计算相似度时不考虑词汇之间的顺序.

给定两组不同语言的实体描述文本 d_s 和 d_t,其中 d_s来自源语言而 d_t来自目标语言,最优对齐就是找到最大化公式(2)的对齐词语组合:

$$\mathrm{optAlign}(d_s,d_t)=\max_{\{(x_s^i,x_t^i)\}_{i=1}^{N}}\sum_{i=1}^{N}\mathrm{wordsim}(x_s^i,x_t^i) \tag{2}$$

其中 d_s 和 d_t 之间词语的最优对齐组合为 $f=\{(x_s^i,x_t^i)\}_{i=1}^{N}$,$x_s^i$ 来自 d_s而 x_t^i 来自 d_t,$\mathrm{wordsim}(x_s^i,x_t^i)$为两个词语的相似度,$N$ 为对齐词语对的数量,该值等于 d_s和 d_t 中较短的文本长度. 式(2)可以用 Kuhn-Munkres[20]算法求解. 由于文本中不同的词语具有不同的重要程度,因此使用 tf-idf[21]权重对(2)进行修正:

$$\mathrm{sim}(d_s,d_t)=0.5\cdot\sum_{(x_s^i,x_t^i)\in f}\alpha_{st}^i\cdot\mathrm{wordsim}(x_s^i,x_t^i)\cdot(w_s^i+w_t^i) \tag{3}$$

其中 i 为 f 中的第 i 组实体对,w_s^i 为词语 x_s^i 的 *tf-idf* 值,α_{st}^i和 $\mathrm{wordsim}(x_s^i,x_t^i)$ 的值根据词语类型的不同有不同的计算方式. 如下图 2 所示,相同实体的不同语言描述可能会出现一些完全相同的词语,例如表示年份的数字和人名等. 因此对于完全相同的词语 x_s^i 和 x_t^i,令 $\mathrm{wordsim}(x_s^i,x_t^i)=1$;当 x_s^i 和 x_t^i 分别具有词向量 $\boldsymbol{x}_s^i$ 和 $\boldsymbol{x}_t^i$ 时,可令 $\mathrm{wordsim}(x_s^i,x_t^i)=\cos(\boldsymbol{W}\boldsymbol{x}_s^i,\boldsymbol{x}_t^i)$;而当 x_s^i 或 x_t^i 没有对应的词向量时 $\mathrm{wordsim}(x_s^i,x_t^i)=0$(训练词向量时要求词语在语料中的最低词频为 5,因此部分词语没有词向量). α_{st}^i的值通常为 1,而当 d_s和 d_t之间具有例如数字和人名等相同的词语时它们描述同一实体的概率会增加,此时对 α_{st}^i赋予一个大于 1 的值.

实体Douglas Dean Smail对应的英语版本描述

Douglas Dean Smail(born 2 September 1957 in Moose Jaw, Saskatchewan) is a retired professional ice hockey left winger who played in the NHL for thirteen seasons from 1980 through 1993.

实体Douglas Dean Smail 对应的德语版本描述

Douglas Dean Smail(* 2. September 1957 in Moose Jaw, Saskatchewan) ist ein ehemaliger kanadischer Eishockeyspieler, der in seiner aktiven Zeit von1975 bis 1996 unter anderem für die Winnipeg Jets,Minnesota North Stars, Nordiques de Québec und Ottawa Senators in der National Hockey League spielte

图2　一组对齐实体的描述信息

4.2　基于共享参数模型的跨语言知识向量

本文采用 TransE 模型训练知识向量,该模型假设关系为实体之间的翻译,即对任意三元组(h,r,t)都期望满足 $\boldsymbol{t}=\boldsymbol{h}+\boldsymbol{r}$,其对应的能量函数为:

$$f(h,r,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\| \tag{4}$$

对任一语言 $i\in I$,其对应 KG_i的目标函数为:

$$G_i=\sum_{(h,r,t)\in \mathrm{KG}_i,(h',r',t')\in \mathrm{KG}_i^-}[\gamma+f(h,r,t)-f(h',r',t')]_+ \tag{5}$$

其中$[x]_+ = \max\{0, x\}$，KG_i^- 为 KG_i 的负采样[12]集合. 对于参与跨语言实体对齐的 KG_{i_1} 和 KG_{i_2}，总的目标函数为：

$$G = G_{i_1} + G_{i_2} \tag{6}$$

为了实现实体对齐需要将不同语言的知识向量集成到共同的语义空间. 本文采用的方法是基于 ILLs 并结合共享参数模型[16]. 共享参数模型使不同语言 KG 的已对齐实体共享相同的词向量，即对于任一已对齐实体对 $(e_1, e_2) \in \mathrm{ILL}(i_1, i_2)$ 有：

$$\boldsymbol{e}_1 \equiv \boldsymbol{e}_2 \tag{7}$$

由于共享参数模型没有正则化变量，因此没有分数函数.

4.3 迭代对齐

在跨语言知识向量的训练过程中，各语言知识向量共享空间的调节作用使可以被对齐的实体的知识向量逐渐接近. 通常判断实体是否对齐的做法是设置阈值 σ_1，当两个实体的知识向量相似度超过 σ_1 即可以判定它们对齐. 本文采用余弦测度衡量知识向量相似度. 对于源语言的一个未对齐实体，在目标语言中可能有多个实体和该实体的知识向量相似度都超过 σ_1，此时仅将相似度最大的两个实体对齐可能会导致判断错误. 为了克服这个问题，本文基于实体的描述信息利用 OA-CDS 模型辅助判断实体是否对齐. 如果源语言 KG_{i_1} 中的一个未对齐实体 e_1 有描述信息且在目标语言 KG_{i_2} 中与该实体的知识向量相似度值大于 σ_1 的未对齐实体也有描述信息，则可使用 OA-CDS 模型计算它们之间描述信息的相似度并设置阈值 σ_2，在文本相似度值大于 σ_2 的实体对之间选择值最大的一组作为最终的对齐实体. 为了保证最终选择出的实体相对其它实体具有足够大的区分度，本文又设置了一个参数 gap，只有文本相似度排第一名的描述大于阈值 σ_2 并且高于第二名的差值大于 gap，该实体对才会被选择.

文献[18]中的实验表明新对齐的实体中如果有错误会导致误差传播. 为了缓解误差传播的影响，本文采用了两种策略，一是迭代对齐时将知识向量重新初始化，二是软对齐.

4.3.1 重新初始化(RE, Reinitialize)

在每轮迭代中，当跨语言知识向量的训练结果稳定后，在知识向量相似度值大于 σ_1 的未对齐实体对中使用 OA-CDS 模型基于实体描述信息选择出新对齐实体对集合 $\mathrm{ILL}_{\mathrm{new}}$，此时将知识向量重新初始化开始新一轮的训练. 在新一轮的训练中，当知识向量的训练稳定后需要将上一轮的 $\mathrm{ILL}_{\mathrm{new}}$ 清空再重新生成新的 $\mathrm{ILL}_{\mathrm{new}}$.

4.3.2 软对齐(SA, Soft Alignment)

对于 $\mathrm{ILL}_{\mathrm{new}}$ 中的任意实体对 (e_1, e_2)，本文参照文献[16]提出的软对齐方法定义了如下的分数函数：

$$K = \sum_{(e_1, e_2) \in \partial_{\mathrm{new}}} (S(e_1, e_2) + S(e_2, e_1)),$$

$$S(e_1, e_2) = \sum_{(e_1, r, t)} f(e_2, r, t) + \sum_{(h, r, e_1)} f(h, r, e_2), \tag{8}$$

对于 KG_{i_1} 和 KG_{i_2}，$\mathrm{ILL}_{\mathrm{new}}$ 中的实体对不会加入共享参数模型的 $\mathrm{ILL}(i_1, i_2)$ 中而是仅仅使用软对齐的方法参与训练，因此在每轮训练中模型总的目标函数为：

$$L = G + K \tag{9}$$

5 实验

本文主要将所提模型应用于跨语言实体对齐任务. 此外，本文还测试了 OA-CDS 模型的性能和跨语言实体对齐任务对 ILLs 数量的敏感性.

数据集 本文基于 Dbpedia 构建了用于进行跨语言实体对齐的数据集(DS-CEA, Date Set for Cross-lingual Entity Alignment)，该数据集包括英语到法语和英语到德语两部分. 本文首先从 Dbpedia(2016-10)中英语到法语(En-Fr)和英语到德语(En-De)的 ILLs(Dbpedia)中各随机采样了 1000 个词语对，要求词语对中的每个词语都在 mapingbased_objects(Dbpedia 中提供三元组的文件)中至少出现 5 次，并限定每个实体的类型都为 Person；再基于两组词语对从各语言的 mapingbased_objects 中选择和词语对中的词语匹配的三元组，最终在这些三元组里还可以利用 ILLs(Dbpedia)再匹配出一部分词语对(例如之前选出的 1000 对词语对中包括词语对 *A-B*，它们匹配出了三元组(A, $r1$, C)和(B, $r2$, D)，虽然 C 和 D 未在之前的 1000 个词语对中出现，但是 *C-D* 存在于 Dbpedia 提供的 ILLs 中)，这些词语对和之前采样的 1000 个词语对构成每个数据集的 ILLs(DS-CEA).

实体的描述信息从 Dbpedia 提供的 short_abstract 里提取，在本数据集中 80% 以上的实体都有描述信息. 在训练的过程中，每个数据集的 ILLs 都按照 30%，10% 和 40% 的比例划分成了训练集、验证集和测试集，剩余 20% 的数据用来进行实验 5.3. 关于数据集的各项统计数据如表 1 所示，其中#表示对应项目的数量.

表 1 DS-CEA 的统计数据

数据集		#实体	#关系	#三元组	#描述	比例	#ILLs
En-Fr	English	44731	246	47328	36571	81.76%	6432
	French	16847	193	18835	14246	84.56%	
En-De	English	28549	236	31362	23913	83.76%	6432
	German	23056	87	27532	19084	82.77%	

为了测试本文所提 OA-CDS 模型的性能，本文还创建了一个文本数据集(DS-CDS). 由于本文提出的 OA-CDS 模型针对知识向量相似度接近的实体对应的描述，这些实体对应的描述信息讨论的主题可能比较接

近. 为了创建和跨语言实体对齐模型应用环境相近的数据集,我们首先训练跨语言实体对齐模型. 训练结束后对于任意$(e_1,e_2)\in \mathrm{ILL}(i_1,i_2)$(DS-CEA),首先在源语言的知识向量空间中抽取相似度和 $e1$ 接近的实体所对应的描述 5 条,再在目标语言的知识向量空间中抽取相似度和 e_2 接近的实体所对应的描述 5 条,加上 e_1 和 e_2 所对应的描述信息共 12 条构成一组数据(凑不够的组舍弃),并针对 En-Fr 和 En-De 各抽取了 500 组.

训练　本文采用 CBOW 模型针对各语言的维基百科语料分别训练了 300 维的词向量并使用文献[19]提供的双语词典训练跨语言映射 W. 在跨语言实体对齐的每一轮训练中,知识向量都基于一个截断的正态分布进行初始化,并通过实验确定知识向量的最优维度为 75 维. 本文采用 AdaGrad 算法[22]对目标函数进行优化,在训练的过程中始终保持对知识向量的归一化. 本文将数据集划分为训练集,验证集和测试集,并使用早停(early stop)[23]的方法使实验终止.

5.1　跨语言实体描述相似度模型性能

本文将 OA-CDS 模型在数据集 DS-CDS 上测试性能. OA-CDS 基于可以从两个方向进行训练的跨语言词向量,因此分别计算英语作为源语言到目标语言法语/德语的的相似度以及法语/德语作为源语言到目标语言英语的相似度.

本文将 OA-CDS 未改进前的源模型(OA-CDS(origin))和贪婪关联算法(GA-CDS)[20]作为基准算法. OA-CDS(origin)主要没有考虑实体描述中的特征;GA-CDS 的主要思想是给源语言描述中的每个词在目标语言的描述中找到和其最相似的词.

实验的过程就是计算 DS-CDS 中每组数据源语言中已对齐实体所对应的文本和目标语言的每个文本的相似度并按从大到小的顺序排序. 该实验所使用的指标为 Hits@1,即正确的描述在所有组当中排名第一所占的比例. 实验的结果如表 2 所示:

表 2　跨语言实体描述相似度模型性能

方向	En→De	De→En	En→Fr	Fr→En
指标	Hits@1	Hits@1	Hits@1	Hits@1
OA-CDS	93.7	93.1	92.4	92.8
OA-CDS(origin)	75.4	76.5	77.8	78.6
GA-CDS	63.9	65.2	58.6	60.3

实验结果表明与 OA-CDS(origin)相比本文所提算法有较大提高,说明本文所提算法可有效利用实体描述中有利于实体对齐的信息. 而 GA-CDS 的效果最差,说明基于最优对齐模型(OA-CDS 和 OA-CDS(origin))的算法相对更加有效.

同时可以看出,总体上从相对稀疏的语料法语、德语映射到英语比从相反的方向进行效果稍好. 可能的原因是将英语的词向量投射到法语、德语的词向量空间后,由于法语、德语的稀疏性更难找到语义接近的词语.

5.2　跨语言实体对齐

本文采用文献[18]中使用的两种指标衡量本文所提模型的性能. (1)Hits@k:正确的实体在 top k 个实体中所占比例;(2)Mean:正确实体的平均排序值. 本文采用的指标是 Hits@1 和 Hits@10. 更高的 Hits@k 和更低的 Mean 表示更好的实验效果.

本文采用的基准算法有线性映射 LM(Linear Mapping)、MtransE 等. 同跨语言词向量类似,LM 方法就是采用式(1)在已经独立训练好的各语言知识向量之间学习映射. MtransE 分为 MtransE-LT(Linear Transformation model),MtransE-TB(Translation Based model)和 MtransE-AC(Distance-based Axis Calibration),其中 MtransE-AC 算法与共享参数模型思想接近但复杂度更高,因此本文采用 MtransE-LT 和 MtransE-TB 作为基准算法.

为了验证 RE 和 SA 策略的作用,本文分别采用文献[16]中的硬对齐(HA,Hard Alignment)方法、单独使用 SA 的方法、RE 和 HA 策略相结合的方法作为基准算法. 为了验证共享参数模型(Ps-TransE)的作用,将其替换成 MtransE-LT 作为基准算法.

如表 3 所示,本文所提方法取得了最佳效果,而 LM 是所有方法中效果最差的. 可能的原因是在 LM 方法独立学习不同语言的知识向量,而其它方法在学习知识向量的过程中都考虑了不同知识向量空间之间的关联从而具有更好的相关性.

Ps-TransE 模型和 MtransE 模型在没有考虑迭代对齐的情况下在本数据集上效果相近;在仅使用 HA 策略的情况下,二者都有小幅提升;在仅考虑 SA 策略的情况下,Ps-TransE 模型比仅考虑 HA 时有了进一步的提升,而 MtransE 模型的效果与没有考虑迭代对齐时相比几乎没有提升,说明 SA 方法在本数据集上不适应于 MtransE 模型,可能的原因是 SA 策略应用到 MtransE 模型时因为参数过多(MtransE 模型有转换矩阵)导致了过拟合;在同时考虑 RE 和迭代对齐的情况下,Ps-TransE 模型的效果比没有考虑迭代对齐和仅考虑迭代对齐时取得了较大幅度的提升,并且软对齐(RE + SA)的效果好于硬对齐(RE + HA). MtransE 模型在使用硬对齐(RE + HA)后相比没有考虑迭代对齐和仅考虑迭代对齐时也取得了较大幅度的提高,而 MtransE 模型在使用软对齐(RE + SA)后相比没有考虑迭代对齐和仅考虑迭代对齐时仅取得了小幅度的提高.

从上图还可以看出,从相对稀疏的语言向英语对齐时比反方向的总体效果要好,可能的原因是从英语的知识向量空间投射到法语、德语的知识向量空间之后,由于法语、德语的相对稀疏性更难找到匹配的实体.

表 3 跨语言实体对齐模型性能

方向	En→De			De→En			En→Fr			Fr→En		
指标	Hits@1	Hits@10	Mean	Hits@1	Hits@10	Mean	Hits@1	Hits@10	Mean	Hits@1	Hits@10	Mean
LM	0.0634	0.1755	2028.9	0.0684	0.1853	1984.3	0.0328	0.1430	2637.4	0.0362	0.1476	2517.3
MtransE-LT	0.1732	0.3589	867.2	0.1711	0.3641	816.7	0.1457	0.3369	1024.4	0.1528	0.3386	1043.1
MtransE-TB	0.0823	0.2646	1531.4	0.0927	0.2703	1544.0	0.0545	0.2197	1872.7	0.0523	0.2268	1757.5
Ps-TransE	0.1678	0.3624	906.2	0.1784	0.3721	887.1	0.1422	0.3413	1082.6	0.1471	0.3436	1057.2
MtransE-LT(RE+SA)	0.1762	0.3610	828.6	0.1768	0.3673	821.5	0.1511	0.3421	1004.4	0.1543	0.3395	984.6
MtransE-LT(RE+HA)	0.2539	0.4660	717.5	0.2651	0.4734	778.8	0.2083	0.3967	767.4	0.2118	0.4092	752.1
MtransE-LT(SA)	0.1739	0.3566	849.1	0.1726	0.3662	827.6	0.1485	0.3353	1028.1	0.1536	0.3364	1018.5
MtransE-LT(HA)	0.1787	0.3622	832.8	0.1756	0.3723	814.4	0.1548	0.3421	986.0	0.1551	0.3439	977.4
Ps-TransE(RE+SA)	**0.3324**	**0.5419**	**622.4**	**0.3373**	**0.5461**	**615.3**	**0.2768**	**0.4526**	**742.1**	**0.2811**	**0.4581**	**716.3**
Ps-TransE(RE+HA)	0.3164	0.5326	684.3	0.3218	0.5346	706.2	0.2535	0.4381	761.2	0.2680	0.4472	741.9
Ps-TransE(SA)	0.1974	0.3857	831.1	0.2039	0.4013	782.4	0.1586	0.3610	974.1	0.1631	0.3647	953.8
Ps-TransE(HA)	0.1721	0.3709	885.2	0.1826	0.3891	873.3	0.1489	0.3527	1032.5	0.1528	0.3542	993.5

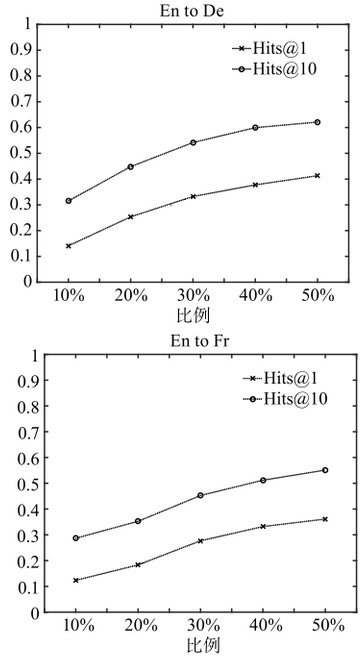

图3 跨语言实体对齐任务对ILLs比例的敏感性

5.3 跨语言实体对齐任务对 ILLs 比例的敏感性

保留数据集中固定 10% 和 40% 的 ILLs 作为验证集和测试集,而分别取 10% 至 50% 的数据作为训练集. 随着训练集在 ILLs 中所占比例的变化,跨语言实体任务的性能如图 3 所示.

实验结果表明实验的性能会随着训练集在 ILLs 中占比的增加而逐渐提升,说明实体对齐的关键就是找到更多的有效 ILLs. 本文所提方法就是通过迭代对齐逐渐找到更多的 ILLs,因此即使在仅提供较少的 ILLs 的情况下任务也能取得一定的效果.

6 结论

本文提出了一种结合知识图谱的内部结构信息和实体描述信息共同进行跨语言实体对齐的模型. 该模型首先通过训练基于共享参数模型的知识向量找到可能被对齐的实体对,再结合实体描述信息利用共享参数模型选出最终的对齐实体,最后通过基于重新初始化和软对齐的策略重复前两个步骤找到更多的对齐实体直到训练结束. 实验结果表明,与基准算法相比,本文所提模型在跨语言实体对齐任务上可以取得相对不错的结果.

参考文献

[1] Zhang Z,Zhuang F,Qu M,et al. Knowledge graph embedding with hierarchical relation structure[A]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing[C]. Brussels, Belgium: Association for Computational Linguistics,2018. 3198 – 3207.

[2] Paulheim H. Data-Drivenjoint debugging of the DBpedia mappings and ontology[A]. European Semantic Web Conference[C]. Vienna,Austria:Springer,2017. 404 – 418.

[3] Miller G A. WordNet:a lexical database for English[J]. Communications of the ACM,1995,38(11):39 – 41.

[4] Rebele T,Suchanek F,Hoffart J,et al. YAGO:Amultilingual knowledge base from Wikipedia, Wordnet, and Geonames [A]. International Semantic Web Conference[C]. Kobe,Japan:Springer,2016. 177 – 185.

[5] Zhang Y, Dai H, Kozareva Z, et al. Variational reasoning for question answering with knowledge graph[A]. Thirty-Second AAAI Conference on Artificial Intelligence[C]. New Orleans,USA:AAAI Press,2018. 6069 – 6076.

[6] Zaremoodi P,Buntine W L,Haffari G,et al. Adaptiveknowledge sharing in multi-task learning:improving low-resource neural machine translation[A]. Meeting of the Association for Computational Linguistics[C]. Melbourne, Australia: Association for Computational Linguistics,2018. 656 – 661.

[7] Faria D,Pesquita C,Santos E,et al. The agreementmakerlight ontology matching system[A]. OTM Confederated Interna-

tional Conferences" On the Move to Meaningful Internet Systems"[C]. Berlin,Germany:Springer,2013. 527 – 541.
[8] Tigrine A N,Bellahsene Z,Todorov K. Light-weight cross-lingual ontology matching with LYAM + +[A] OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"[C]. Rhodes, Greece: Springer,2015. 527 – 544.
[9] Bordes A,Weston J,Collobert R,et al. Learning structured embeddings of knowledge bases[A]. National Conference on Artificial Intelligence[C]. San Francisco, USA: AAAI Press,2011. 301 – 306.
[10] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [A]. Neural Information Processing Systems[C]. Lake Tahoe,USA:Curran Associates Inc,2013. 926 – 934.
[11] Bordes A, Usunier N, Garciaduran A, et al. Translating embeddings for modeling multi-relational data[A]. Neural Information Processing Systems[C]. Lake Tahoe, USA:Curran Associates Inc,2013. 2787 – 2795.
[12] Wang Z,Zhang J,Feng J,et al. Knowledge graph embedding by translating on hyperplanes[A]. National Conference on Artificial Intelligence[C]. Quebec City,Canada: AAAI Press,2014. 1112 – 1119.
[13] Ji G,Liu K,He S,et al. Knowledge graph completion with adaptive sparse transfer matrix[A]. National Conference on Artificial Intelligence [C]. Phoenix, USA: AAAI Press,2016. 985 – 991.
[14] Guo S,Wang Q,Wang L,et al. Jointly embedding knowledge graphs and logical rules[A]. Empirical Methods in Natural Language Processing[C]. Austin,Texas,USA:Association for Computational Linguistics,2016. 192 – 202.
[15] Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment [A]. International Joint Conference on Artificial Intelligence [C]. Melbourne,Australia:AAAI Press,2017. 1511 – 1517.
[16] Zhu H,Xie R,Liu Z,et al. Iterative entity alignment via joint knowledge embeddings[A]. International Joint Conference on Artificial Intelligence[C]. Melbourne,Australia:AAAI Press,2017. 4258 – 4264.
[17] Sun Z,Hu W,Zhang Q,et al. Bootstrapping entity alignment with knowledge graph embedding[A]. International Joint Conference on Artificial Intelligence [C]. Stockholm,Sweden:AAAI Press,2018. 4396 – 4402.
[18] Sun Z,Hu W,Li C,et al. Cross-lingual entity alignment via joint attribute-preserving embedding[A]. International Semantic Web Conference 2017 [C]. Vienna, Austria: Springer,2017. 628 – 644.
[19] Glavas G,Francosalvador M,Ponzetto S P,et al. A resource-light method for cross-lingual semantic textual similarity [J]. Knowledge Based Systems,2017,143(7):1 – 9.
[20] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly,1955,2(1):83 – 97.
[21] Zuo Y,Li J,Tang Y,et al. A value classification of electronic product reviews based on maximum entropy[J]. Chinese Journal of Electronics,2016,25(6):1071 – 1078.
[22] 宋攀,景丽萍. 基于神经网络探究标签依赖关系的多标签分类[J]. 计算机研究与发展,2018,55(8):1751 – 1759.
Song Pan,Jing Liping. Exploiting label relationships in multi-label classification with neural networks[J]. Journal of Computer Research and Development,2018,55(8):1751 – 1759. (in Chinese)
[23] Blanchard G,Hoffmann M,Reis M,et al. Early stopping for statistical inverse problems via truncated SVD estimation[J]. Electronic Journal of Statistics,2018,12(2): 3204 – 3231.

作者简介

康世泽(通讯作者)　男,1991 年生,内蒙古呼伦贝人. 战略支援部队信息工程大学博士研究生,研究方向为知识图谱. E-mail:xiaozebixia@163.com

吉立新　男,1970 生,江苏淮安人. 战略支援部队信息工程大学研究员、博士生导师,研究方向为电信网信息关防.

刘树新　男,1987 生,山东潍坊人. 战略支援部队信息工程大学助理研究员,研究方向为复杂网络、移动通信网安全.

丁悦航　女,1995 年生,山东济南人. 战略支援部队信息工程大学硕士研究生,研究方向为知识图谱.