实体对齐研究综述

张富 杨琳艳 李健伟 程经纬

(东北大学计算机科学与工程学院沈阳 110169)

摘 要 实体对齐(Entity Alignment)旨在发现不同知识图谱(Knowledge Graph)中指代相同事物的实体，是知识图谱融合的关键技术，近年来受到了广泛的关注。早期，研究者们使用字符串的各种特征来进行实体对齐工作。近年来，随着知识表示学习(Knowledge Representation Learning)技术的不断发展，研究者们提出了许多基于知识表示学习的实体对齐方法，效果明显优于传统方法。然而，实体对齐的研究仍然存在着许多亟待解决的问题与挑战，比如数据质量、计算效率等。

本文从实体对齐的定义、数据集和评价指标出发，详细深入地综述和比较了传统实体对齐方法和基于知识表示学习的实体对齐方法。针对传统方法，分类介绍了基于相似性计算和基于关系推理的实体对齐方法，并深入研究了每类方法对字符特征、属性特征、关系特征的利用，同时深入分析了不同方法之间的优势与不足。针对基于知识表示学习的实体对齐方法，本文进行了重点讨论、分析和对比。首先，本文将该类实体对齐方法抽象为由三个模块(即嵌入模块、交互模块和对齐模块)组成的统一框架，依据三个模块对每个方法进行了详细的综述。进一步地，根据方法所利用的信息种类的不同，将已有方法划分为基于结构信息、属性信息、实体名信息、实体描述信息和综合信息等八类方法，对每一类方法进行了详细的综述。然后，对基于知识表示学习的实体对齐方法进行了深入对比分析。最后，讨论了实体对齐工作的主要挑战，包括稀疏知识图谱的处理、标注数据的缺乏和噪声问题、方法的效率问题等，并对该工作的未来进行了展望。

关键词 知识图谱；实体对齐；知识图谱融合

中图法分类号 TP18

An Overview of Entity Alignment Methods

ZHANG Fu YANG Lin-Yan LI Jian-Wei CHENG Jing-Wei

(School of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning, 110169)

Abstract Entity alignment, which aims at finding entities that refer to the same thing in different knowledge graphs, is a key step of knowledge graph fusion. The traditional methods used various characteristics of strings to align entities. In recent years, with the continuous development of knowledge representation learning technology, researchers have proposed many entity alignment methods based on knowledge representation learning, the effect is obviously better than the traditional methods. However, there are still many problems and challenges in the research of entity alignment, such as data quality, computational efficiency and so on.

Starting from the concept, datasets, and evaluation criteria of entity alignment, this paper makes a detailed and in-depth overview about traditional entity alignment methods and entity alignment methods based on knowledge representation learning. The traditional entity alignment methods are classified into the methods based on similarity calculation and based on relation reasoning. The character feature, attribute feature and relation feature in each method are investigated in detail, and also the advantages and disadvantages of different

本课题得到国家自然科学基金面上项目(No.61672139)资助.**张富**，博士，副教授，博士生导师，CCF会员（55285M），主要研究领域为知识图谱、语义Web.E-mail: zhangfu@cse.neu.edu.cn.**杨琳艳**，博士研究生，主要研究领域为知识图谱实体对齐.E-mail: 1712085416@qq.com.**李健伟**，硕士研究生，主要研究领域为知识图谱实体对齐.E-mail: 418442508@qq.com.**程经纬(通信作者)**，博士，讲师，硕士生导师，CCF会员（22767M），主要研究领域为知识图谱、语义Web.E-mail: chengjingwei@cse.neu.edu.cn.

methods are analyzed and compared. In particular, the paper emphatically discusses the entity alignment methods based on knowledge representation learning, and abstracts these entity alignment methods into a unified framework, which includes three main modules: embedded module, interaction module, and alignment module. Also, the methods are further classified into eight categories according to the structure information, attribute information, entity name information, entity description information, and integrated information. Further, this paper makes a deep comparison and analysis of these methods. Finally, the main challenges and further directions of entity alignment are discussed and summarized, including the processing of sparse knowledge graph, the lack and noise problems of annotation data, and the efficiency of the method.

Key words Knowledge Graph; Entity Alignment; Knowledge Graph Fusion

1 引言

近几年，互联网的快速发展促使各领域建立了越来越多包含互补信息的大规模知识图谱(Knowledge Graph)。同时，随着链接数据(Linked Data)[1]计划的发展，网络上语义数据的数量不断增加，而各应用领域面临的主要挑战之一就是集成越来越多独立设计且存在于不同知识图谱中的实体，使得大规模的知识图谱之间可以高效协调。因此，如何发现不同知识图谱实例之间的链接成为各个领域亟待解决的重要问题[1]。

尤其是，随着近几年知识图谱的快速发展，涌现出大量的知识图谱[2]。然而，目前很多的知识图谱由不同机构和个人构建，这些知识图谱的需求特定，设计和构建并不统一，因此互相之间存在异构和冗余问题。知识融合旨在将知识图谱中的异构和冗余等信息进行对齐和合并，形成全局统一的知识标识和关联[1]。实体对齐(Entity Alignment，EA)[3, 4]是知识图谱融合过程的关键技术，主要目的是发现不同知识图谱之间的等价实体。由于不同知识图谱的知识内容存在来源各异和人为理解不同，指代同一个事物的文字表达会各有不同。这是不同知识图谱融合集成的显著问题，影响共享数据的实现。因此，针对基于知识图谱的知识融合研究，对后续大数据集成统一的技术探索和发展意义重大[5]。

实体对齐一般可以分为本体对齐和实例对齐，本体对齐重点关注类、属性和关系，而实例对齐则更加注重真实世界中指代的具体事物[2]。早期的相关工作主要集中在本体对齐方面，近几年随着机器学习和深度学习的发展，也逐渐向实例对齐方向发展。本体对齐相对于实例对齐而言更加笼统概括，主要针对包含相似实例的一类实体；而实例对齐对信息的精细程度要求更多，也更加复杂。此外，实体对齐任务与传统的实体消歧(链接)任务存在差异，传统的实体消歧需要将文本内容中提及的实体，链接到知识图谱或知识图谱中的实体。然而实体对齐，是将两个或者多个结构化的知识图谱或知识图谱中的实体进行等价对齐[6]。

随着实体对齐技术的发展，许多学者提出了不同种类的实体对齐方法，涌现出大量的实体对齐研究文献。早期，研究者们使用字符串的各种特征来进行实体对齐工作。近些年，随着知识表示学习(Knowledge Representation Learning)技术的快速发展，研究者们提出了许多基于知识表示学习的实体对齐方法，这些方法取得了比传统方法更好的效果。然而，截止目前仍然缺少有关实体对齐技术全面而深入的方法综述。已有的综述文献[7]主要概括了传统实体对齐方法；文献[8]仅针对基于图神经网络(Graph Neural Network，GNN)的实体对齐方法进行了简略介绍；文献[5]和[9]从实验的角度，对部分实体对齐方法在数据集上的性能进行了深入比较分析。与上述已有综述不同，本文从方法和技术层面，更加全面深入地综述和比较了传统实体对齐方法和基于知识表示学习的实体对齐方法，对这些已有方法进行了详细的划分与综述。针对传统方法，本文深入分析研究了每类方法对字符特征、属性特征、关系特征的利用，进而对比了不同方法之间的优势与不足。针对主流的基于知识表示学习的实体对齐方法，本文深入挖掘并研究了每种方法所利用的知识图谱信息，根据所利用信息种类的不同将已有方法细分为八个类别，同时进行了详细的综述和对比分析。

基于以上分析，本文将实体对齐方法分为两大类，一类是传统的实体对齐方法，一类是基于知识

[1]https://lod-cloud.net

表示学习的实体对齐方法。在给出实体对齐的问题定义、数据集和评价指标的基础上，进一步详细深入地综述和比较了这两大类方法。主要贡献如下：

- 针对传统方法，分类介绍了基于相似性计算和基于关系推理的实体对齐方法，并深入研究了每类方法对字符特征、属性特征、关系特征的利用，同时深入分析了不同方法之间的优势与不足。
- 针对基于知识表示学习的实体对齐方法，本文进行了重点讨论、分析和对比：（*i*）本文将该类实体对齐方法抽象为由三个模块(即嵌入模块、交互模块和对齐模块)组成的统一框架，依据三个模块对每个方法进行了详细的综述；（*ii*）根据方法所利用的知识图谱信息种类的不同，将已有方法细分为基于结构信息、属性信息、实体名信息、实体描述信息和综合信息等八类方法，并对每类方法进行了详细介绍和分析；（*iii*）进一步对基于知识表示学习的实体对齐方法进行了深入对比分析。分析结果表明，科学有效的迭代方法和对多种信息的利用都能够提升方法的性能等。
- 讨论了实体对齐工作的主要挑战和未来方向，包括稀疏知识图谱的处理、标注数据的缺乏和噪声问题、方法的效率问题等。

本文后续章节安排如下：第 2 节给出实体对齐的问题定义、数据集和评价指标；第 3 节介绍传统实体对齐方法；第 4 节综述基于知识表示学习的实体对齐方法；第 5 节概括实体对齐工作的主要挑战和未来方向；最后给出本文总结。

2 问题定义、数据集和评价指标

2.1 问题定义

知识图谱 $G = (E, R, T)$是一个有向图，其中包括实体的集合 E、关系的集合 R 和三元组的集合 $T \subseteq E \times R \times E$[1]。给定源知识图谱 $G_1 = (E_1, R_1, T_1)$、目标知识图谱 $G_2 = (E_2, R_2, T_2)$以及已对齐实体对(训练集)$S = \{(u, v) \mid u \in E_1, v \in E_2, u \equiv v\}$，其中≡代表等价，即实体 u 和实体 v 指向的是同一个事物，实体对齐任务的目标就是发现这两个知识图谱中等价的实体对。如图 1 所示，实体对齐任务的目标是发现等价实体对。

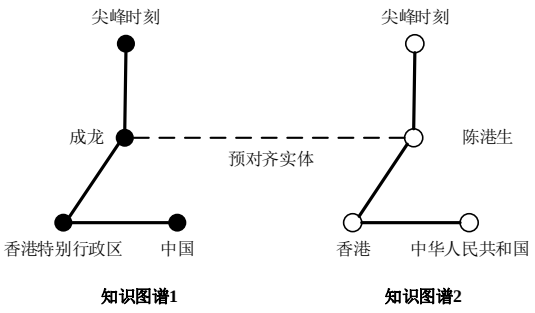


图 1 实体对齐任务示意图

2.2 数据集

传统的实体对齐主要集中在本体对齐，所以传统方法的性能测评主要是基于本体对齐评估计划(OAEI, Ontology Alignment Evaluation Initiative)**Error! Reference source not found.**所提供的评测系统。此外，部分数据集也是由互联网数据构建而成。传统实体对齐方法主要的数据集如表 1 所示。

基于知识表示学习的实体对齐数据集的构建方法主要分为两种：基于已有的知识图谱和基于互联网信息构建的数据集。下面 2.2.1 和 2.2.2 节分别对这两种数据集进行详细介绍。

2.2.1 基于已有知识图谱构建的数据集

许多实体对齐数据集都是基于 DBpedia[2]、Wikidata[3]、YAGO[4]等知识图谱构建的。DBpedia 是一个从多种语言版本的 Wikipedia 中提取结构化知识的大型知识图谱。YAGO 是由德国马普所构建的包含了来自 Wikipedia、WordNet[5]、GeoNames[6](免费的全球地理数据库)等多个不同数据源的多语言知识图谱。

文献[18]使用 DBpedia 不同语言版本(英语-汉语、英语-日语、英语-法语)构建了实体对齐数据集 DBP15K，包含三个子版本数据集 DBP15KZH-EN、DBP15KJA-EN、DBP15KFR-EN，如表 2 所示。

此外，部分方法使用的数据集是基于多个知识图谱构建而成。文献[19]建立的实体对齐数据集是基于 DBpedia、LinkedGeoData(一个来源于公开地图 OpenStreetMap 的空间知识图谱)、GeoNames 和 YAGO 构建而成，该文献将 DBpedia 与其它三个知

[2]https://wiki.dbpedia.org/
[3]https://www.wikidata.org/wiki/Wikidata:Main_Page
[4]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago

[5]https://wordnet.princeton.edu/
[6]http://www.geonames.org/

识图谱分别进行实体对齐，数据集的详情如表 3 所　　示。

表 1 传统实体对齐数据集

文献	数据来源	数据集	实体数
Cohen 等人(2002)[11]	Cora projects	Cora	1916
	Organization names	OrgName	116
	Restaurant guide	Restaurant	864
	National park name lists	Parks	646
Sarawagi 等人(2002)[12]	CiteSeer	Bibliography data	254
	Local telephone company of Pune in India	Address data	300
Jean-Mary 等人(2009)[13]	OAEI(2008)	OAEI(2008)	-
Arasu 等人(2010)[14]	Internet organization records	QRG	2×10^6(records)
	Publication domain	PUB	1.1×10^6 (records)
Suchanek 等人(2011)[15]	YAGO	YAGO	2795289
	DBpedia	DBpedia	2365777
	IMDb	IMDb	4842323
Lacoste 等人(2013)[16]	YAGO	YAGO	1.4M
	IMDb	IMDb	3.1M
	Freebase	Freebase	474K
Song 等人(2016)[17]	ACM、DBLP、CiteSeer、EPrints、IEEE、Newcastle、ECS	RKB	8.2×10^7(三元组)
	CiteSeer、DBLP	SWAT	2.6×10^7(三元组)

表 2 DBP15K[18]数据集详情

数据来源	数据集		实体数	关系数	属性数	关系三元组	属性三元组
DBpedia	DBP15K	Chinese	66,469	2,830	8,113	153,929	379,684
		English	98,125	2,317	7,173	237,674	567,755
		Japanese	65,744	2,043	5,882	164,373	354,619
		English	95,680	2,096	6,066	233,319	497,230
		French	66,858	1,379	4,547	192,191	528,665
		English	105,889	2,209	6,422	278,590	576,543

文献 **Error! Reference source not found.**使用的数据集是基于 FB15K 和 DBpedia 的英语版、法语版、德语版构建而成。其中 FB15K 是由 Freebase 建立的一个稠密的知识图谱，通常在知识表示学习任务中作为基准。该文献基于上述两个知识图谱构建了三个数据集：En-Fr、En-De 和 Fb-Db。其中 En-Fr 是基于 DBpedia 的英语版和法语版，En-De 是基于 DBpedia 的英语版和德语版，Fb-Db 是基于 FB15K 和 DBpedia 的英语版。

表 3 DBP-LGD、DBP-GEO、DBP-YAGO[19]数据集详情

数据来源	数据集		实体数	关系三元组	属性三元组
DBpedia(DBP),	DBP-LGD	LGD	24,309	10,084	90,054

LinkedGeoData(LGD), GeoNames(GEO), YAGO	DBP-GEO	DBP	22,748	19,594	166,008
		GEO	21,794	17,410	98,790
	DBP-YAGO	DBP	22,748	19,594	166,008
		YAGO	30,628	38,451	173,309
		DBP	33,627	36,906	184,672

文献[21]指出，上述已有实体对齐数据集的分布较为稠密，而真实世界中知识图谱的分布往往是稀疏的。为此，文献[21]提出了四个数据集，分别是 EN-FR、EN-DE、DBP-WD、DBP-YG，这四个数据集来源于 DBpedia、Wikidata 和 YAGO3。文献[22]使用了该数据集测试了所提出的基于重排序的迭代式实体对齐方法。

文献[23]也基于 DBpedia、Wikidata 和 YAGO3 构建了两个实体对齐数据集 DBP-WD 和 DBP-YG，每个数据集都包含 100k 个已对齐的实体对。

2.2.2 基于互联网信息构建的数据集

部分实体对齐数据集是基于互联网中的信息构建而成，包括网站中的信息、一些公开的数据集等。

文献[24]和[4]构建了两个数据集，第一个数据集来自于英文的 Cora[7]，记为 Cora1。Cora 记录的是科技论文的书目信息，每个实体包含作者、题目等属性。文献[24]和[4]将含义相同的两个实体分别分配到两组中，独特的实体被随机分配到两组其中的一组中，这些独特的实体被看作是噪声实体。第二个数据集来自于中文的百度和豆瓣影视信息，记为 Baidu_Douban_M/TV。文献[25]使用的数据集的来源也是 Cora、百度和豆瓣影视信息。文献[26]根据 Wikidata、Freebase、IMDB[8]和 Amazon Music[9]，构建了 Movie Dataset 和 Music Dataset。基于互联网信息构建的实体对齐数据集如表 4 所示。

[7] https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz
[8] https://www.imdb.com/
[9] https://music.amazon.com/

表 4 基于互联网信息构建的实体对齐数据集

数据集	来源	包含	文献
Cora	科学论文引文中提取的科学论文书目信息	作者、标题、出版商	SEEA(2019)[24]、融合语义和结构信息的实体对齐方法(2019)[4]、自适应属性选择的实体对齐方法(2020)[27]、基于重排序的迭代式实体对齐(2020)[22]
百度	百度网站电影条目	名称、演员、导演、类型、发布时间	SEEA(2019)[24]、融合语义和结构信息的实体对齐方法(2019)[4]、自适应属性选择的实体对齐方法(2020)[27]、基于重排序的迭代式实体对齐(2020)[22]
豆瓣	豆瓣网站电影条目	名称、演员、导演、类型、发布时间	SEEA(2019)[24]、融合语义和结构信息的实体对齐方法(2019)[4]、自适应属性选择的实体对齐方法(2020)[27]、基于重排序的迭代式实体对齐(2020)[22]
Movie Dataset	IMDB 和 Freebase	电影名称、人物、特征、类型	CG-MuAlign(2020)[26]
Music Dataset	Amazon Music 和 Wikidata	歌曲名称、专辑、演唱者	CG-MuAlign(2020)[26]

2.3 评价指标

实体对齐方法使用的评价指标主要分为两类，接下来对这两类评价指标分别进行介绍。

2.3.1 Hits@k, MR, MRR

Hits@k是指结果排名前k个中存在正确实体的情况所占的比例，Hits@k 越大方法的效果越好。

MR(Mean Rank)代表的是正确对齐实体排名的平均值，MR 越小方法的效果越好。

MRR(Mean Reciprocal Rank)代表的是正确对齐实体排名的倒数的平均值，MRR 越大方法的效果越好。

已有的实体对齐方法使用这三个指标中的一个或多个对结果进行评价。例如，文献 **Error! Reference source not found.**使用 Hits@k 进行评价；文献[18], [19],[28]和[29]使用 Hits@k 和 MR 进行评价；文献[22]和[30]使用 Hits@k 和 MRR 进行评价；文献[23]和[31]使用 Hits@k、MR 和 MRR 进行评价。

2.3.2Precision, Recall, F1-measure

精确率(Precision)表示对齐结果的准确程度，定义为公式(1)所示：

$$Precision = \frac{N_{success}}{N_{total}} \tag{1}$$

其中 N_{success} 表示算法对齐的正确实体对数量，N_{total} 表示算法对齐的实体对总数。

召回率(Recall)的定义如公式(2)所示：

$$Recall = \frac{N_{success}}{R_{total}} \tag{2}$$

其中 N_{success} 表示算法对齐的正确实体对数量，R_{total} 表示所有真实存在的关系数量。

F1 值(F1-measure)用于综合反映整体效果，定义如公式(3)所示：

$$F1 = \frac{2 * Recall * Precision}{(Recall + Precision)} \tag{3}$$

例如，文献[25], [4]和[24]等使用了上述三个指标进行评价。

3 传统实体对齐方法

传统的实体对齐方法大多数都集中在句法和结构上，尤其是早期的实体对齐和映射技术主要侧重于计算实体之间标签和字符的距离。传统的实体对齐方法主要从两个角度解决实体对齐问题：一类是基于相似度计算来比较实体的符号特征[11]，另一类是基于关系推理[32]，最近的研究还使用统计机器学习来提高准确性。**本节将详细综述已有的传统实体对齐方法，同时深入研究每类方法对字符特征、属性特征、关系特征的利用，并进行对比分析。**

3.1 基于相似性计算的实体对齐方法

基于相似度计算来进行实体对齐的方法，主要

利用词频-逆文档频率(TFIDF, Term Frequency-Inverse Document Frequency)[11]，主动学习和机器学习分类器以及NGram匹配/编辑距离/数字匹配等[12]，同义词集和语义验证[13]，以及过滤机[14]等技术计算实体之间的相似性，实际上就是在相似性计算的基础上加入了机器学习分类器[12]、主动学习[12], [14]、语义验证[13]以及过滤阻塞[14]等技术来提高实体对齐算法的性能。

3.1.1 基于 TFIDF 的实体对齐方法

TFIDF是统计学方法的一种，常用于信息检索。词频(TF, Term Frequency)代表了一个词语在文档中出现的次数，通常进行归一化处理。如公式(4)所示：

$$TF_i = \frac{词语i出现的次数}{文件中的总词数} \tag{4}$$

逆文档频率(IDF, Inverse Document Frequency)的核心思想是若包含某个词语的条目越少，则这个词语具有很好的条目区别性，则 IDF 的值越大。计算如公式(5)：

$$IDF_i = log\left(\frac{文件中条目总数}{包含词语i的条目数+1}\right) \tag{5}$$

TFIDF 的计算如公式(6)所示：

$$TFIDF_i = TF_i * IDF_i \tag{6}$$

Cohen等人[11]使用TFIDF来计算实体名称之间的距离，距离在阈值范围则表示二者匹配。通过遍历所有实体对$E_1 \times E_2$来获得候选实体对集，训练二分类函数来进行实体对齐。算法示意图如图2所示。

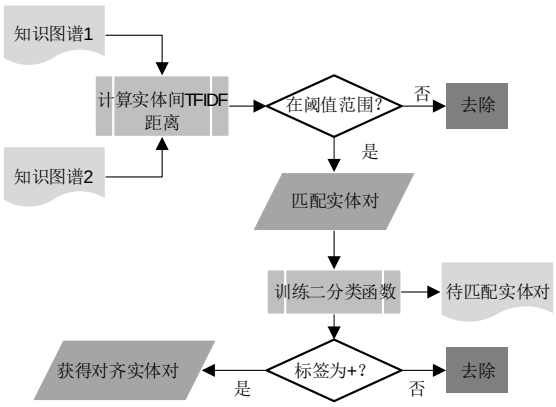


图 2 TFIDF 实体对齐示意图

算法的目标是获得一组匹配实体对集$y=\{(e_1, e_1^{'}), (e_2, e_2^{'}), \ldots, (e_n, e_n^{'})\}$，$e_1$和$e_1^{'}$分别是$E_1$和$E_2$中的实体，且$y \subseteq E_1 \times E_2$。$y^* = \{(e_1^*, e_1^{'*}), (e_2^*, e_2^{'*}), \ldots (e_n^*, e_n^{'*})\}$表示一组已对齐的正确实体对集，损失函数定义如公式(7)：

$$Loss \equiv |\{(e_k, e_k^{'}) \in y : (e_k, e_k^{'}) \notin y^*\}| + |\{(e_k^*, e_k^{'*}) \in y^* : (e_k^*, e_k^{'*}) \notin y\}| \tag{7}$$

3.1.2 基于机器学习和主动学习的实体对齐方法

Sarawagi 等人[12]指出在构建实体对齐数据集时，由于利用人工来手动发现覆盖全面的实体匹配对非常困难，而且人工筛选费时费力，限制了算法的可扩展性。

为此，Sarawagi 等人[12]提出了基于主动学习的方法，使得系统能够交互发现机器难以识别的实体对，从而提升了系统的性能。此外，Sarawagi 等人认为实体对的匹配问题可以看作分类问题，即匹配、非匹配和模糊匹配。因此，使用机器学习构造了三个分类器来进行实体对齐，分别是决策树(Decision Tree)[33]、朴素贝叶斯(Naive Bayesian)[34]和支持向量机(SVM, Support Vector Machine)[35]。针对三个分类器的训练，为了加速训练同时提高精度，设立了委员会机制。假设三个分类器的分类结果一致，则作为匹配实体对；若结果出现差异，则使用人工交互的方式对此类实体对进行标记。使用机器学习和主动学习的示意图如图3所示。

再者，对于大型知识图谱，实体对$E_1 \times E_2$的大小可能会超出控制范围，考虑到算法的可扩展性，Sarawagi 等人[12]在候选匹配实体对集的生成过程中进行了简单的过滤操作(即当实体对的首字母相同时，进行匹配操作)，并通过对数据进行采样来减少整体数据集的大小。同时在计算生成的候选匹配实体的相似性时，进一步利用了实体的文本属性(包括 NGram 匹配、重叠单词分数、编辑距离)、数字匹配、以及空匹配填充等技术。

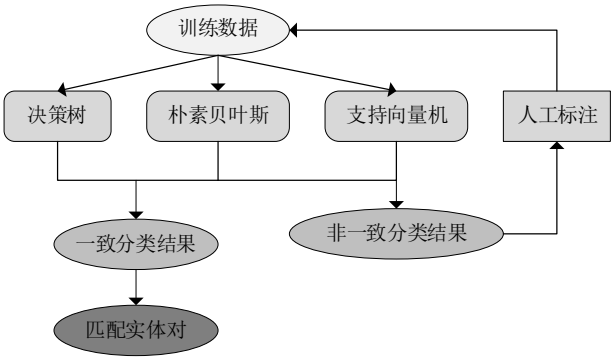


图 3 基于机器学习和主动学习进行实体对齐示意图

3.1.3 基于同义词集和语义验证的实体对齐方法

Jean-Mary 等人[13]提出了 ASMOV(Automated Semantic Matching of Ontologies with Verification)算法，该算法在实体词法特征上使用额外的同义词典进行相似性计算，同时使用语义验证进一步对实体语义加以考虑。

对于两个知识图谱中的实体对(e,e')，每个实体包含的标签、ID 和注释信息进行词法特征相似性S_l计算，假设额外的词典集为Γ，同义词集为$syn(w)$，反义词集为$ant(w)$。词法特征相似性计算S_l如公式(8)所示：

$$S_l(e,e')=\begin{cases}1.0, & if\ w=w'\\ 0.99, & if\ w'\in syn(w)\\ 0.0, & if\ w'\in ant(w)\\ Lin(w,w'), & if\ w\in \Gamma\wedge w'\wedge\Gamma\wedge w'\notin syn(w)\\ \dfrac{tok(w)\cap tok(w')}{max(|tok(w)|,|tok(w')|)}, & otherwise\end{cases} \tag{8}$$

其中$Lin(w,w')$表示由 Lin 等人提出的信息相似性计算方法[36]，$tok(w)$代表 w 中的有序字符串，$|tok(w)|$表示有序字符串的数量。进一步地，大于相似性阈值的实体对需要再次通过语义验证来确定是否存在语义矛盾，即通过制定多实体对齐、交叉对齐、不相交假设等语义规则，进一步严格处理已对齐实体，从而得到更精准的对齐实体。

3.1.4 基于过滤机制和主动学习的实体对齐方法

Arasu 等人[14]认为以往的主动学习方法无法保证用于训练的数据质量，并且没有标准接口，从而得到的结果有很强的不可预知性。其次，由于遍历实体集$E_1\times E_2$时，输入扩展性差，无法处理大型知识图谱。再者，他们指出实体对齐问题可以简单地看作是二分类问题，即匹配和非匹配。但实体对齐与分类问题的主要区别在于：数据集中非匹配的数据量远远超过匹配的数据量。如何在候选匹配实体中平衡匹配数据与非匹配数据显得尤为重要。为此，Arasu 等人[14]提出了结合过滤机制和主动学习的实体对齐方法。

Arasu 等人[14]提出将二分类函数$\mathcal{B}$同实体分类函数$\mathcal{M}$相结合的方式，进行端到端训练，即对输入实体对先进行二分类函数筛选，通过的实体对再进行详细计算。将过滤机制同主动学习结合，可以减少对标记数据的需求。

过滤机制使用的二分类函数如公式(9)所示：

$$\mathcal{B}: E_1\times E_2\longrightarrow true, false \tag{9}$$

对于知识图谱实体对齐过程中只需要对$\mathcal{B}(e_1,e_2)=true$的实体对进行后续详细计算。两个候选实体进行相似性计算时，利用实体的属性信息计算 Jaccard 距离[37]。如公式(10)和(11)所示：

$$\mathcal{J}=\frac{|e_1\cap e_2|}{|e_1\cup e_2|}\# \tag{10}$$

$$d_{\mathcal{J}}(e_1,e_2)=1-\mathcal{J}(e_1,e_2) \tag{11}$$

3.2 基于关系推理的实体对齐方法

基于关系推理的实体对齐方法，主要利用了知识图谱中实体之间的关系，通过构造概率函数[15]、关系相似函数[16]、关系可比性函数[17]来推理关系之间的语义等价性，进而实现关系相应的实体之间的对齐。

3.2.1 基于概率函数的实体对齐方法

Suchanek 等人[15]提出了 PARIS(Probabilistic Alignment of Relations, Instances, and Schema)算法，该算法通过将关系进行函数化，进一步基于关系推理原则进行概率计算，将实体对齐概率化，进而实现实体的对齐任务。

首先对实体之间的关系进行函数化，假设两个实体之间的关系用三元组来表示：(e_1,r,e_2)，则存在函数关系(x,r,y)，且定义$fun(r,x)$为：

$$fun(r,x)=\frac{1}{|y:r(x,y)|} \tag{12}$$

其中$|y:r(x,y)|$表示当头实体 x 确定时满足关系 r 的尾实体的数量。然后，定义关系逆函数为：

$$fun^{-1}(r,x)=fun(r^{-1},x) \tag{13}$$

最后定义关系函数$fun(r)$：

$$fun(r)=\frac{|x:\exists y:r(x,y)|}{|x,y:r(x,y)|} \tag{14}$$

在此基础上，根据公式(15)中的关系推理公式，进一步将实体匹配概率化，也就是计算两个实体可能匹配的概率。

$$\exists r,y,y^{'}:r(x,y)\wedge r(x^{'},y^{'})\wedge y\equiv y^{'}\wedge fun^{-1}(r)\ is\ high\implies x\equiv x^{'} \tag{15}$$

将上面的推理公式概率化可以得到如公式(16)所示的实体可能匹配的概率公式$Pr_1(x\equiv x^{'})$：

$$Pr_1(x\equiv x^{'}):=1-\prod_{r(x,y),r(x^{'},y^{'})}\left(1-fun^{-1}(r)\times Pr(y\equiv y^{'})\right) \tag{16}$$

其中概率Pr_1表示两个实体可能匹配的概率。由于在实际推理过程中同样存在不匹配的情况，所以需要进一步定义两个实体不匹配的概率Pr_2，相应的关系推理公式如(17)所示：

$$\exists r,y:(x,y)\wedge(\forall y^{'}:r(x^{'},y^{'})\wedge fun(r)\ is\ high\implies y\not\equiv y^{'})\implies x\not\equiv x^{'} \tag{17}$$

将上面推理公式概率化可以得到两个实体不匹配的概率公式(18)：

$$Pr_2(x\equiv x^{'}):=\prod_{r(x,y)}\left(1-fun(r)\prod_{r(x^{'},y^{'})}\left(1-Pr(y\equiv y^{'})\right)\right) \tag{18}$$

最后综合将两个概率公式(16)和(18)合并得到最终的实体对齐概率Pr_3，如公式(19)所示：

$$Pr_3(x\equiv x^{'}):=Pr_1(x\equiv x^{'})\times Pr_2(x\equiv x^{'}) \tag{19}$$

通过最后的概率值Pr_3来确定两个实体是否对齐。

3.2.2 基于关系相似函数的实体对齐方法

Lacoste 等人[16]提出了 SiGMa(Simple Greedy Matching for Aligning LargeKnowledge Bases)算法，该算法采用贪心思想，利用实体的字符串、属性和结构信息进行局部搜索来完成不同知识图谱之间的实体对齐。实体对齐过程主要分为两个部分：(*i*)构造基于属性信息的相似性函数；(*ii*)利用已匹配的实体对构建邻接图。SiGMa 算法利用了实体的属性信息和关系信息，对属性三元组和关系三元组进行计算，来获取对齐实体。

假设对于两个知识图谱的实体集$E_1\times E_2$，用矩阵y表示，则$y_{ij}=(i,j):i\in E_1,j\in E_2$，若两个实体 i、j 为匹配实体，则$y_{ij}=1$，否则$y_{ij}=0$。在实体匹配过程中需要计算实体相似性S_{ij}和关系相似性G_{ij}，构造实体匹配目标函数$obj(y)$如下公式(20)：

$$obj(y)\doteq\sum_{(i,j)\in E_1\times E_2}y_{ij}\left[(1-\alpha)S_{ij}+\alpha G_{ij}(y)\right]\quad where\ \ G_{ij}\doteq\sum_{(k,l)\in\mathcal{N}_{ij}}y_{kl}\,w_{ij,kl} \tag{20}$$

其中α是在实体相似性和关系相似性之间的平衡参数，实验中设置为 0.25；$\mathcal{N}_{ij}$表示实体 i 和 j 的局部邻接实体集；$w_{ij,kl}$表示已知实体 k 和 l 匹配的情况，是计算关系相似性G_{ij}的权重参数，需要根据邻接实体的贡献程度确定权重$w_{ij,kl}$，计算公式如(21)所示：

$$w_{ij,kl}=\gamma_i w_{ik}+\gamma_j w_{jl} \tag{21}$$

其中γ_i和γ_j是正则化参数，假设$\mathcal{N}_i$表示实体 i 在知识图谱G_1中的邻接实体集，则正则化参数的计算如公式(22)所示：

$$\gamma_i\doteq\frac{1}{2}\left(1+\sum_{k\in\mathcal{N}_i}w_{ik}\right)^{-1}\qquad\gamma_j\doteq\frac{1}{2}\left(1+\sum_{l\in\mathcal{N}_j}w_{jl}\right)^{-1} \tag{22}$$

实体的相似性S_{ij}的计算分为两部分：实体及属性字符串相似性$string(i,j)$和属性集相似性$prob(i,j)$，公式如(23)所示：

$$S_{ij}=(1-\beta)string(i,j)+\beta prob(i,j) \tag{23}$$

其中β是平衡参数，实验中设置为 0.25。假设实体e_1包含属性$p_1,p_2,\dots,p_{n_1}$，且属性值为$v_1,v_2,\dots,v_{n_1}$，实体e_2包含属性$q_1,q_2,\dots,q_{n_2}$，且属性值为$l_1,l_2,\dots,l_{n_2}$。字符串相似性$string(i,j)$和属性集相似性$prob(i,j)$计算公式如(24)所示：

$$string(i,j) = \frac{\sum_{v\in(\mathcal{W}_i\cap\mathcal{W}_j)}(w_v^1+w_v^2)}{smoothing+\sum_{v\in\mathcal{W}_i}w_v^1+\sum_{v'\in\mathcal{W}_j}w_{v'}^2} \quad (24)$$

$$prob(i,j) = \frac{\sum_{(a,b)\in M_{12}}(w_{p_a,v_a}^1+w_{q_b,l_b}^2)Sim_{p_a,q_b}(v_a,l_b)}{2+\sum_{a=1}^{n_1}w_{p_a,v_a}^1+\sum_{b=1}^{n_2}w_{q_b,l_b}^2}$$

其中$\mathcal{W}_i$表示实体 i 中的词语集，w_v^1表示词语 v 在实体集E_1中的逆文件频率 IDF(见 3.1.1 节)，$smoothing$是加在分母中的平滑项，使得长字符串中尽可能地包含更多的共现短单词。M_{12}代表已对齐的属性集，w_{p_a,v_a}^1表示实体e_1属性值的 IDF，w_{q_b,l_b}^2表示实体e_2属性值的 IDF，$Sim_{p_a,q_b}(v_a,l_b)$是一个[0,1]函数，若$v_a=l_b$则为 1，否则为 0。

3.2.3 基于关系可比性函数的实体对齐方法

Song 等人[17]提出了两种可扩展的实体对齐算法 HistSim(Candidate Selection Based On Instance Matching History)和 DisNGram(Candidate Selection Using DiscriminatingPredicate N-Grams)。HistSim 主要依靠一组启发式方法执行筛选，从而去除匹配可能性很小的实例对；DisNGram 不依赖于任何实际的实体匹配算法来筛选不匹配的实例对，它以无监督的方式学习候选实体对，并采用索引技术可伸缩地选择类似的实例对，并且通过附加更细粒度的字符级相似性度量来生成最终候选实例对。

对于一个实体序列$\{e_1,e_2,\dots,e_m\}$，HistSim 从第一个e_1开始依次同后面的实体进行对比，若两实体相似则加入相似实体集合$H(e_1)$。为了避免实体对重复比较，每一个实体只同其后面实体进行匹配计算。若两实体的字符串前缀匹配，并且余弦相似性大于指定阈值，则加入相似实体集合，$H(e_1)$中的实体根据相似性进行降序排序。对比两实体(e_1,e_2)的相似性时，计算$HSim\big(H(e_1),H(e_2)\big)$，如公式(25)所示。若两实体$(e_1,e_2)$的相似实体存在很大程度的相似性(通过构建 *sigmoid* 函数动态控制阈值)，则得到实体对齐结果。

$$HSim\big(H(e_1),H(e_2)\big) = \begin{cases} +\infty, & H(e_1)=\emptyset \ or \ H(e_2)=\emptyset \\ \frac{|H(e_1)\cap H(e_2)|}{|H(e_1)\cup H(e_2)|}, & otherwise \end{cases} \quad (25)$$

DisNGram 算法主要分为两部分：(1)学习具有区分性的关系谓词。(2)根据学习的关系选择可能匹配的实例。

(1) 学习具有区分性的关系谓词

知识图谱中的关系三元组(e_1,r,e_2)，需要计算关系 r 的可区分性$dis(r)$和覆盖性$cov(r)$。假设C_{e_1}表示实体e_1的同类实例组合，则可区分性$dis(r)$和覆盖性$cov(r)$的计算如下公式(26)和(27)所示：

$$dis\big(r,C_{e_1},G\big) = \frac{|\{e_2|t=<e_1,r,e_2>\in G\wedge e_1\in C_{e_1}\}|}{\big|t\big|t=<e_1,r,e_2>\in G\wedge e_1\in C_{e_1}\big|} \quad (26)$$

$$cov\big(r,C_{e_1},G\big) = \frac{|\{e_1|t=<e_1,r,e_2>\in G\wedge e_1\in C_{e_1}\}|}{\big|C_{e_1}\big|} \quad (27)$$

(2) 选择可能匹配的实例

通过学习到的具有区分性的关系谓词 r，选择此关系的三元组 t，对实例三元组进行字符级比较。给定两个不同三元组(t,t')，字符级比较计算如下公式(28)所示：

$$S(t,t') = \frac{|ngram(t)\cap ngram(t')|}{min(|ngram(t)|,|ngram(t')|)} \quad (28)$$

为了进一步对实例进行比较，对实例的属性信息进行了相似性Sim和比率$Ratio$计算。对于两个属性(p_1,p_2)的计算如公式(29)和(30)：

$$Sim(p_1,p_2) = \frac{|tokenset(p_1)\cap tokenset(p_2)|}{min(|tokenset(p_1),tokenset(p_2)|)} \quad (29)$$

$$Ratio(p_1,p_2) = \frac{min(|tokenset(p_1),tokenset(p_2)|)}{max(|tokenset(p_1),tokenset(p_2)|)} \quad (30)$$

其中$tokenset(p_1)$表示三元组中所有以属性p_1为谓词的主语（也就是三元组中的 objects）的集合。

3.3 传统实体对齐方法的对比分析

正如上文所述，传统实体对齐方法主要集中在本体对齐，因此在实体对齐任务上也没有统一的性能比较。表 5 对传统实体对齐方法进行了对比分析。

从上述方法可以看出，由于实体对齐过程中，实体的各种属性不同以及涉及的领域也不同，很难给出统一的相似度计算函数。同时这种离散的属性信息忽略了多方面隐含的语义信息(例如属性之间的关联语义以及三元组结构之间的语义信息等)，使

得对齐效果有限。因此，近年来，越来越多的学者开始提出新的实体对齐技术(详见第 4 节)。

表 5 传统实体对齐方法对比

	文献	构造候选实体	特征计算	优势	不足
基于相似性计算	Cohen 等人 (2002)[11]	$E_1 \times E_2$	TFIDF，字符串，前缀	复杂性低，易计算	输入扩展性差
	Sarawagi 等人 (2002)[12]	$E_1 \times E_2$首字母和采样过滤	机器学习+主动学习	对输入数据集简单过滤，提升了对大规模数据的处理能力	人工标注种子实体：匹配和非匹配
	Jean-Mary 等人 (2009)[13]	$E_1 \times E_2$	标签、ID、注释+同义词集+语义验证	多种相似性计算，提升实体对齐的精度	计算复杂，依赖实体大量的额外信息，对大型数据处理性能差
	Arasu 等人 (2010)[14]	$E_1 \times E_2$阻塞函数过滤	属性集+机器学习+主动学习+过滤机制	集成阻塞函数和匹配函数，减少对标记数据的需求，运行速度加快	只考虑实体的属性信息，未加入结构信息
	文献	**构造候选实体**	**关系利用**	**优势**	**不足**
基于关系推理	Suchanek 等人 (2011)[15]	$E_1 \times E_2$	关系等价推理概率化	不仅仅对齐实体，同时对齐关系和类	只能处理一对一的关系
	Lacoste 等人 (2013)[16]	迭代预对齐种子实体的邻接实体	根据实体间关系构建邻接图	利用了实体的属性信息和结构信息	需要预先对齐实体、关系和属性，需要大量标记数据
	Song 等人 (2016)[17]	根据对齐的关系谓词选择	量化关系谓词	不依赖任何相似性计算方法，无监督学习，不需要标注数据	需要提前获取实体类别属性信息，构建候选实体速度慢

4 基于知识表示学习的实体对齐方法

表示学习又叫做表征学习(Representation Learning)，其目的是利用机器学习技术将描述对象表示为低维稠密的向量，两个向量之间的距离反映的是两个对象之间的语义关系。将表示学习应用于知识表示中，即知识表示学习(Knowledge Representation Learning)，目的是实现知识图谱中实体和实体之间关系的向量表示，通过降低高维实体和关系，得到低维向量的数值表示。

基于知识表示学习技术能够将实体和关系表示为低维向量空间的能力，**许多研究者们提出了基于知识表示学习的实体对齐方法，该类方法也成为目前解决实体对齐问题的主要技术**。通过深入研究这些方法，**本文概括并抽象出一个统一的实体对齐框架**，如图 4 所示。其基本思想就是首先通过知识表示学习技术对知识图谱进行嵌入，即**嵌入模块**；之后根据已对齐的实体对将不同知识图谱的嵌入空间映射到同一个向量空间中，即**交互模块**；最后根据向量空间中实体之间的距离或者相似度得到实体对齐结果，即**对齐模块**。此外，大多数方法还引入了迭代机制，将实体对齐结果添加至已经对齐的实体对中。

本节接下来将对基于知识表示学习的实体对齐方法进行重点介绍、对比分析和总结。首先，依据图 4 提到的三个模块(即嵌入模块、交互模块和对齐模块)对每一种方法进行了详细介绍。同时，本文通过深入研究，对所有方法根据其利用的知识图谱信息的不同进行了详细的分类(见 4.2 节)。然后进一步对该类方法进行了详细的对比，并对结果进行了深入的分析(见 4.3 节)。

下面 4.1 节首先简单介绍现有的知识表示学习技术，然后后续几节重点综述基于知识表示学习的实体对齐方法，并进行深入的对比分析。

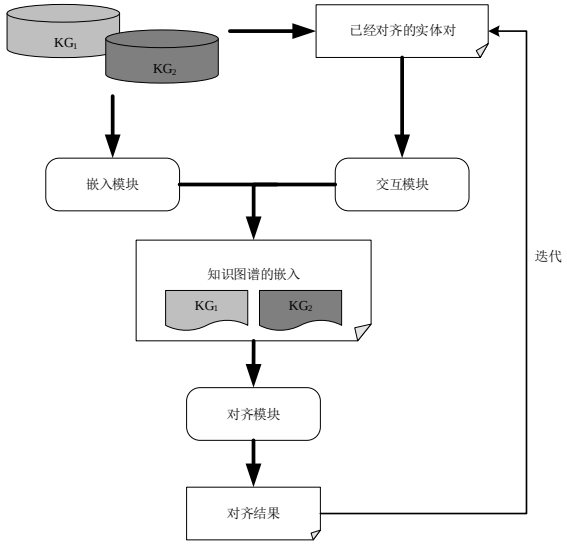


图 4 基于知识表示学习的实体对齐方法基本框架

4.1 知识表示学习技术

目前主要的知识表示学习技术可以分为三类：翻译模型、语义匹配模型、深度模型[38], [39]。

4.1.1 翻译模型

基于 word2vec 的词向量模型，Mikolov 等人发现训练得到的词向量在词向量空间进行平移操作之后，可以保持不变[40], [41]。例如 C(king) - C(queen) ≈C(man) - C(woman)，其中的 $C(w)$表示 word2vec 模型学习到的单词 w 的词向量，研究者们受此现象启发提出了多种翻译模型。具有代表性的翻译模型包括 TransE[42]及其扩展模型，这类模型将关系视为头实体到尾实体的翻译。

在 TransE[42]中，对于每个三元组（h, r, t），将关系看作是头尾实体的连接，并且关系存在方向，TransE 也被叫做翻译模型。具体地，如图 5 所示，对于每个三元组（h, r, t）来说，TransE 希望 $h + r \approx t$。

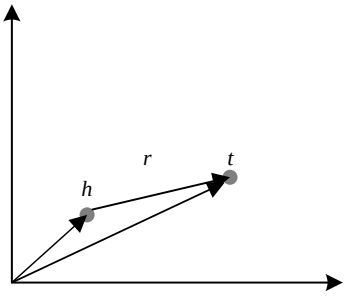


图 5 TransE 模型

在 TransE 中，三元组（h, r, t）的评分函数如下公式(31)，在计算得分时采用 L_1 或 L_2 距离：

$$f_r(h,t) = \|h + r - t\|_{L_1/L_2} \tag{31}$$

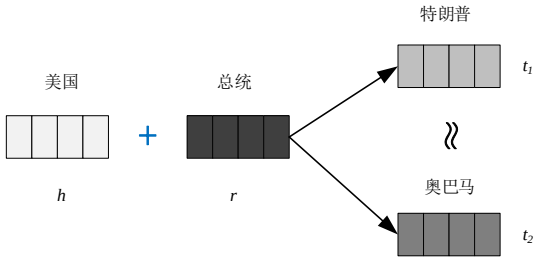


图 6 1-N 复杂关系示例

然而，TransE 仍然存在着诸多局限性，包括：

(1) 处理复杂关系时的局限性，这里的复杂关系指的是 1-N、N-1 和 N-N。图 6 所示的就是一个典型的 1-N 复杂关系，美国的总统会指向多个尾实体，而 TransE 会使得这些尾实体的表示相同，然而事实上，“特朗普”实体和“奥巴马”实体存在着诸多不同，这就降低了 TransE 对实体的区分性。

(2) TransE 孤立学习每个三元组，没有考虑多步关系路径，而关系路径中蕴含着许多有用信息。

针对这些局限性，研究者们提出了许多 TransE 的扩展模型。针对复杂关系处理，TransH[43]让同一个实体在不同关系下的表示不同，TransR[44]让不同的关系拥有不同的语义空间，TransA[45]将评分函数中的距离改为马氏距离，并为每一维学习不同的权重。针对多步关系路径，提出了 PTransE[46]，如图 7 所示，在 PTransE 中，如果一个关系路径起到的作用和一个关系相同，那么就将这个关系路径看做是一个关系。

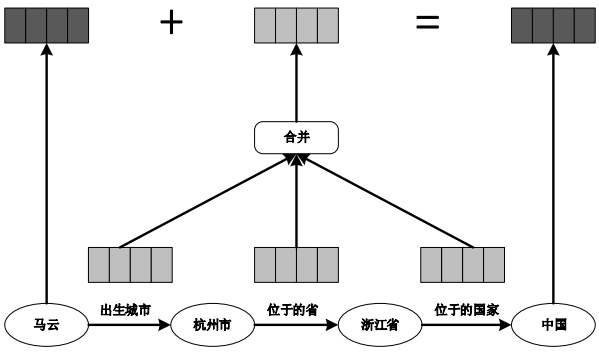


图 7 PTransE 模型

4.1.2 语义匹配模型

语义匹配模型主要使用基于相似度的方法来推断关系事实，包括 DistMult[47]、ComplEx(Complex Embeddings)[48]、HolE(Holographic Embeddings)[49]等。

DistMult[47]采用关系特定的双线性形式来考虑实体和关系之间的联系，三元组（h, r, t）的评分函数如下公式(32)：

$$f_r(h,t) = h^T M_r t \tag{32}$$

其中 M_r 是关系 r 对应的矩阵，为了减少关系参数的数量，DistMult 将 M_r 限制为对角矩阵。

ComplEx[48]在嵌入时考虑了复数值，这使得 ComplEx 能够对各种二元关系进行建模。在 ComplEx 中，三元组(h, r, t)的评分函数如公式(33)：

$$f_r(h,t) = sigmoid(X_{hrt}) \tag{33}$$

当三元组存在时得分应为 1，不存在时得分应为-1。X_{hrt} 的计算方法如下公式(34)：

$$\begin{aligned} X_{hrt} =& < Re(w_r), Re(h), Re(t) \\ & > + \\ & < Re(w_r), Im(h), Im(t) > + \\ & < Im(w_r), Re(h), Im(t) > + \\ & < Im(w_r), Im(h), Re(t) > \end{aligned} \tag{34}$$

其中 w_r 是关系对应的复数向量，$Re(x)$表示 x 的实部，$Im(x)$表示 x 的虚部。

HolE[49]使用向量的循环相关来表示实体对，循环相关*：$R_d \times R_d \to R_d$ 的计算方法如下公式(35)：

$$[a * b]_k = \sum_{i=0}^{d-1} a_i b_{(i+k) \bmod d} \tag{35}$$

在 HolE 中，三元组（h, r, t）的评分函数如公式(36)所示：

$$f_r(h,t) = sigmoid\left(r^T (h * t)\right) \tag{36}$$

4.1.3 深度模型

深度模型主要使用深度学习技术进行知识表示学习，包括 ProjE(Projection Embedding)[50]、ConvE(Convolutional 2D Embeddings)[51]、R-GCN(Relational Graph Convolutional Networks)[52]等(如表 6 所示)。其中 ProjE 利用多层感知机进行建模；ConvE 利用卷积和全连接层对实体和关系的联系进行建模；R-GCN 则利用图卷积网络(GCN)[53]进行建模；ConvKB[54]将实体和关系建模为相同大小的嵌入向量，对每个三元组的嵌入连接到一个输入矩阵；ConvR[55]用不同位数的一维向量表示实体嵌入和关系嵌入，三元组的得分是通过点积结合神经网络输出得到的；CapsE[56]采用胶囊网络(Capsule networks, CapsNets)[57]进行实体和关系建模；RSN[58]使用随机游走方式选择实体并学习三元组的关系路径。

表 6 部分深度模型

深度模型	核心思想	损失函数
ProjE[50]	给定两个输入嵌入 ,将预测任务看作排序问题	成对排名（Rank）
ConvE[51]	二维卷积嵌入链接预测	交叉熵
R-GCN[52]	多关系数据的图卷积神经网络	交叉熵
ConvKB[54]	关系建模、二进制分类器	负对数似然
ConvR[55]	对关系进行充分卷积	交叉熵
CapsE[56]	使用胶囊网络进行建模	负对数似然
RSN[58]	随机游走采样在实体之间传递信息	交叉熵

有关翻译模型、语义匹配模型、深度模型的详细介绍可参见文献[38]和[39]。

4.2 基于知识表示学习的实体对齐方法

基于知识表示学习的实体对齐方法已经成为目前解决实体对齐问题的主要技术，并取得了较好的效果，其中绝大多数方法都使用翻译模型或图神经网络(Graph Neural Network, GNN)[59]进行知识表示学习，因为它们有着较强的鲁棒性和泛化能力。

通过分析可以发现，这些方法具有相似框架（如上文图 4 所示）：首先利用翻译模型、GNN 等知识表示学习技术对知识图谱进行嵌入(即嵌入模块)，此时不同知识图谱的嵌入空间是不同的。之后根据已对齐的实体对将不同知识图谱的嵌入空间映射到同一个向量空间中。有些方法通过在数据准备阶段融合不同知识图谱中的元素，进而直接将不同知识图谱映射到同一个向量空间中(即交互模块)。最后根据向量空间中实体之间的距离或者相似度得到实体对齐结果(即对齐模块)。

本文根据每种方法在进行实体对齐时所利用的知识图谱信息的不同，将已有方法细分为八类，如表 7 所示。下面将详细介绍每一类方法，并进行深入的对比分析。

4.2.1 利用结构信息进行实体对齐

该类实体对齐方法借助知识图谱结构信息进

行实体对齐。具体来说，**结构信息**包括：(1)知识图谱中的关系三元组，格式为（E, R, E），例如（中国，首都，北京）。(2)关系路径。(3)邻接信息。

值得注意的是，对实体对齐任务来说，结构信息非常有价值。目前绝大多数实体对齐方法都利用了结构信息。当然，当知识图谱中的结构信息不足时，对齐的效果可能比较有限。

(1) 使用关系三元组作为结构信息的方法

该类方法在关系三元组中捕捉结构特征，使用的知识表示学习方法通常为翻译模型，许多方法在利用结构信息时都采用此策略。

文献[60]提出了一个跨语言知识图谱嵌入模型MTransE，该模型主要包括两个部分：

a) 使用 TransE 模型根据关系三元组分别对两个知识图谱的实体和关系进行嵌入，两个知识图谱的嵌入空间是相互独立的。

b) 构建出两个知识图谱的嵌入空间之间的转换方法，损失函数如下公式(37)：

$$L = \sum_{(T,T')\in\delta(G_1,G_2)} S_a(T,T') \tag{37}$$

其中，G_1 和 G_2 代表的是两个不同语言的知识图谱，$\delta(G_1,G_2)$代表的是两个知识图谱中已经对齐的三元组对组成的集合，$S_a(T,T')$是对已经对齐的三元组对进行评分的方法。该模型提出了三种评分的方法，第一种是基于距离的轴校准，根据两个三元组之间的距离对其进行评分；第二种是平移向量，即将跨知识图谱的对齐也视为一种关系，这种关系也视为翻译的过程；第三种是线性变换，即定义转移矩阵，两个对齐的三元组之间通过转移矩阵进行转换。最后再根据知识图谱的嵌入进行实体对齐，实验结果表明，线性变换的效果最好。

表 7 基于知识表示的实体对齐方法对比

类别	利用的信息	模型	
		模型	**具体利用**
I	结构信息	MTransE[60]	关系三元组
		BootEA[61]	关系三元组+自举训练
		Multi-mappingRelations[62]	关系三元组+路径
		MuGNN[63]	邻接信息+交叉注意力
		KECG[64]	邻接信息+关系权重
		SSP[65]	邻接信息+关系三元组
		TransEdge[66]	关系三元组+关系嵌入
		AliNet[67]	邻接信息+远距离实体
		MRAEA[68]	邻接信息+关系方向
		VR-GCN[30]	邻接信息
		SelfAttention-GCN[69]	邻接信息+距离区分
II	属性信息	Schema-Agnostic[70]	
III	结构信息、属性信息	JAPE[18]	
		CTEA[31]	
		GCN(SE+AE)[71]	
		SEEA[24]	
		COTSAE[72]	
		AttrGNN[73]	
		JarKA[74]	
		AttrE[19]	
		融合语义和结构信息的实体对齐方法[4]	
		自适应属性选择的实体对齐方法[27]	

Ⅳ	结构信息、实体名信息	Cross-KG**Error! Reference source not found.**
		HGCN[75]
		RDGCN[76]
		DGMC[77]
		NMN[78]
		RREA[79]
		RNM[80]
		DAT[81]
		CEA[82]
		基于重排序的迭代式实体对齐[22]
Ⅴ	结构信息、实体描述信息	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[28]
Ⅵ	结构信息、属性信息、实体描述信息	HMAN+BERT[83]
Ⅶ	结构信息、属性信息、实体名信息	MultiKE[23]
		EPEA[84]
Ⅷ	属性信息、实体名信息、实体描述信息、	BERT-INT[85]

文献[86]提出了 AKE(Adversarial Knowledge Embedding)模型，在使用 TransE 进行嵌入过程中加入了对抗学习，使用了 GAN(Generative Adversarial Network)[87]模型进行实体频率采样。AKE 模型如图 8 所示。虚线表示生成器和对抗模块提供的敌对反馈框架包括三个模块，即表示模块、映射模块和对抗模块。表示模块使用两个独立的编码器来学习源知识图谱和目标知识图谱的实体表示；映射模块通过种子信息采用线性变换训练实体对齐；对抗模块维护嵌入分布的对齐。

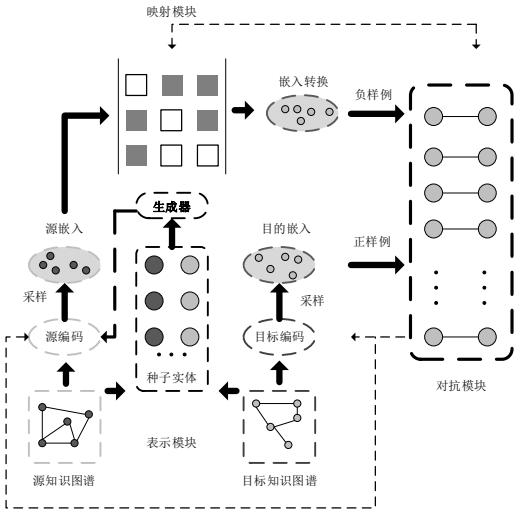


图 8AKE 模型

文献[88]提出了 OTEA(Optimal Transport-based Entity Alignment) 模型，借鉴 Optimal transport(OT)[89]的方法，在实体对齐的交互模块平移矩阵上添加了最优转换的约束。文献[62]将 TransE 和 ComplEx 结合起来进行实体嵌入，能够处理除一对一之外的复杂关系。文献[90]在使用 TransE 进行嵌入表示时，对实体频率进行了区分。文献[66]使用 TransE 进行实体嵌入时，加入了对实体之间关系的嵌入计算。

(2) 使用关系路径作为结构信息的方法

该类方法在关系路径中捕捉结构特征，使用的知识表示学习方法通常也是翻译模型。

文献[29]提出了基于联合知识嵌入和迭代机制的实体对齐算法，算法包括 3 个部分：

a) 分别根据知识图谱中的关系三元组和关系路径使用翻译模型对实体和关系进行嵌入，该算法分别尝试了这两种策略，并进行了对比。对于使用关系三元组进行嵌入的情况，该算法使用的是 TransE；对于使用关系路径进行嵌入的情况，该算法使用的是 PTransE。实验表明，使用 PTransE 时的效果更好。

b) 根据已对齐的实体将不同知识图谱中的嵌入空间映射到同一个语义空间中，该算法提出了三种模型并分别进行了实验。第一种是基于翻译的模型，将实体对齐看作是一种特殊的关系；第二种是线性转移模型，定义一个转移矩阵，对齐的实体之间通过转移矩阵进行转换；第三种是参数共享模型，让对齐的实体拥有相同的嵌入。实验表明，参数共享模型的效果最好。

c) 计算出不同知识图谱的实体之间的语义距离，具体的计算方法与第二步使用的模型有关，根据语义距离进行对齐，之后，使用高置信度的新对齐的实体对集合对联合嵌入进行迭代更新。迭代更新的策略有两种，即直接对齐和软对齐。直接对齐只适用于参数共享模型，直接将高置信度的新对齐的实体对加入到已对齐实体对中。然而，当引入错误对齐时，直接对齐存在错误传播的问题，因此提出了软对齐，为每个高置信度的新对齐的实体对设置一个可靠性评分，软对齐适用于第二步中所有模型。实验表明，软对齐效果更好。

(3) 使用邻接信息作为结构信息的方法

这类方法假设不同知识图谱中表达含义相同的实体具有相似的邻接信息，这类方法大多使用图神经网络(Graph Neural Network, GNN)[59]进行嵌入。

文献[30]使用 VR-GCN(Vectorized Relational Graph Convolutional Network)模型对知识图谱中的实体和关系进行嵌入，该模型综合考虑了 GCN 的优点和知识图谱的平移性质。该算法通过累加相邻实体和关系的嵌入来更新实体的嵌入，通过整合头实体和尾实体的嵌入来更新关系的嵌入。该算法可以应用于神经网络，多层的网络可以捕捉到更灵活的结构信息。算法提出的卷积函数能够区分不同角色的实体，并利用知识图谱的平移性质来学习实体和关系的嵌入，从而有效捕捉了结构信息。之后，该文献进一步提出了基于 VR-GCN 模型的实体对齐框架 AVR-GCN，框架的流程分为两步：

a) 根据已对齐的实体对生成更多的三元组，添加到知识图谱中，从而使得嵌入更加科学，进而提升在对齐任务上的表现。

b) 利用 VR-GCN 模型分别得到两个知识图谱的嵌入，再根据已对齐的实体对计算出转移矩阵，利用转移矩阵将两个知识图谱的嵌入映射到同一个语义空间中，最后根据嵌入之间的距离进行对齐。

与上述文献[30]的思想类似，一些方法也提多了基于 GNN 的实体对齐方法。文献[75]提出 HGCN(Highway-GCN)模型，在 GCN 的基础上添加了高速门机制。文献[76]提出了 RDGCN(Relation-aware Dual-Graph Convolutional Network)模型，使用 GCN 进行嵌入表示，又针对实体之间的关系构建了对偶关系图，提出 GAT(Graph Attention Mechanism)，使用图注意力机制对实体之间的关系分配不同的权重。文献[63]提出 MuGNN(Multi-channel Graph Neural Network)模型，在 GNN 基础上添加了自注意力机制和图交叉注意力机制。文献[64]提出了 KECG(KnowledgeEmbedding model and Cross-Graph model)模型，通过使用投影矩阵约束图注意力机制 GAT 构建交叉图模型，并且结合 TransE 学习知识图谱的向量表示。文献[65]提出 SSP(Structure and SemanticsPreserving)模型，使用文献[75]中的 HGCN 进行全局结构嵌入，并结合 TransE 对邻接实体之间的关系进行表示学习。文献[65]提出 AliNet(KG alignment network)模型，使用 GCN 进行结构信息嵌入，并加入注意力机制选择远距离邻接实体。文献[68]提出 MRAEA(Meta Relation AwareEntity Alignment)模型，使用 GNN 嵌入结构信息，对实体之间的关系进行类型、方向和逆向的区分，采用注意力机制对实体之间的关系分配不同的权重。文献[78]提出 NMN(Neighborhood Matching Network)模型，在 GCN 嵌入时，使用跨图注意力机制来获取邻接实体的差异。文献[91]提出 REA(Robust Entity Alignment)模型，使用 GNN 嵌入，重点加入了噪声数据的处理和检测。文献[77]提出 DGMC(Deep Graph Matching Consensus)模型，使用 GNN 来进行知识图谱局部特征匹配，并提出了同步消息传递迭代来更新实体的邻域信息，将不符合的邻接实体去除。文献[79]提出 RREA(Relational Reflection Entity Alignment)模型，将 GNN 和翻译模型结合，并将关系嵌入和实体嵌入进行结合来完成实体对齐。文献[26]提出 CG-MuAlign(Graph neural network for Multi-type entity Alignment)模型，使用 GNN 对不同类型的实体进行联合对齐，采用节点注意力机制和边注意力机制，对邻接实体和实体间关系进行权重区分。文献[69]提出邻接实体的短距离差异和长距离依赖，使用 GCN 和自注意力机制捕获短距离差异和长距离依赖信息。

4.2.2 利用属性信息进行实体对齐

上述 4.2.1 中的方法使用知识图谱中的结构信息，对不同知识图谱的模式和结构有一定的限制。然而真实世界存在大量异构的知识图谱。因此文献[70]提出了模式无关的实体对齐方法(Schema-Agnostic)。将实体对齐看作是分类问题，生成候选实体对的标签(匹配或非匹配)。将实体的属性值连接，使用 BERT(Bidirectional Encoder Representations from Transformers)进行训练。模型如图 9 所示。

$E = \{e_1, \dots, e_m\}$和$E' = \{e_1', \dots, e_n'\}$表示两个来自不同知识图谱的实体集；通过过滤操作之后获得候选实体对$\{c_1, \dots, c_k\}$，并将其属性连接起来，其中$\{A_1, \dots, A_x\} \in A_E$，$\{A_1', \dots, A_y'\} \in A'_{E'}$。[CLS]表示序列的开始，[SEP]表示序列的结尾，经过 BERT 分类器最后预测标签。

4.2.3 利用结构信息和属性信息进行实体对齐

上述 4.2.1 和 4.2.2 节中的方法分别利用了知识图谱的结构信息和属性信息进行表示学习。仅利用结构信息时，忽略了知识图谱的属性信息，而属性信息具有准确度更高的特点，能够提升实体对齐效果。但是只使用属性信息又忽略了知识图谱的结构和实体之间的联系。为此，一些方法提出了综合利用结构信息和属性信息进行实体对齐。

具体来说，**属性信息**指的是知识图谱中的属性三元组，格式为（*E, A, V*），例如（*John*，年龄，24）。一些实体对齐方法引入了属性信息以补充结构信息，主要分为以下两种。

(1) 利用结构信息和属性的名称

属性相关性嵌入考虑了属性之间的相关性，如果多个属性经常一起用于描述一个实体，那么就认为它们是相关的。例如，经度和纬度相关，因为它们通常组合成为坐标，用于描述一个地点。

文献[18]假设相似实体具有相似相关属性，该算法包括两个模块，分别是根据关系三元组得到的结构嵌入(Structure Embedding, SE)和根据属性三元组得到的属性嵌入(Attribute Embedding, AE)，最后

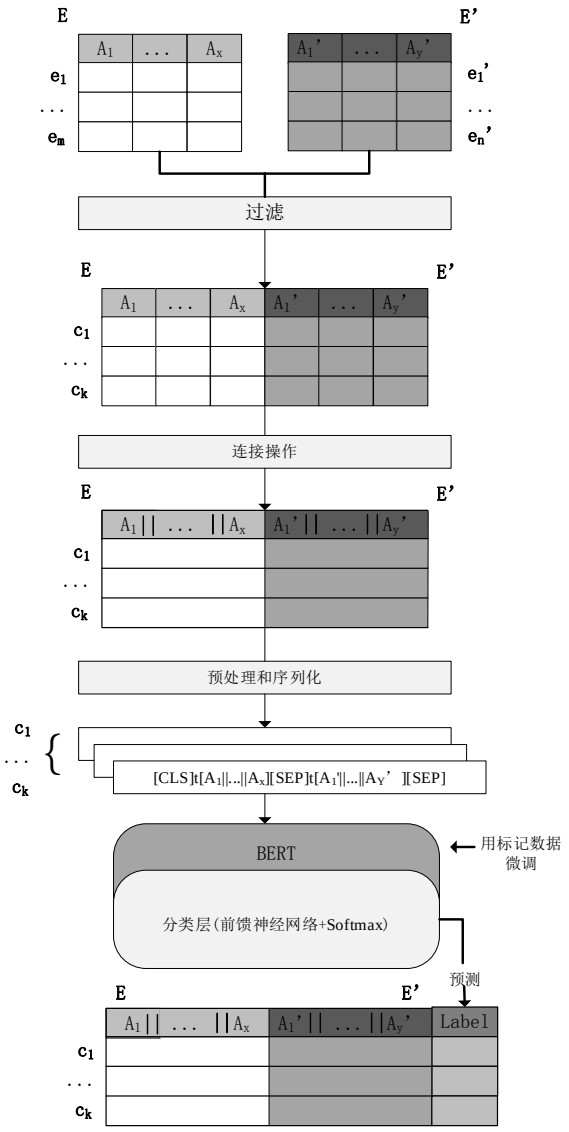


图 9 Schema-Agnostic 模型图

将 SE 和 AE 结合起来得到两个知识图谱中所有实体的嵌入，该算法 JAPE 的框架如图 10 所示。

具体分为以下三个步骤：

a) SE：首先，将已有的关系三元组作为训练的正例，再根据正例生成反例，之后根据正例和反例进行训练，得到实体和关系的嵌入。值得注意的是，在训练时已对齐的实体对共享相同的嵌入。

b) AE：该算法假设已对齐的实体对的属性之间是相关的，且相关的属性的嵌入之间是相近的，根据这两个假设得到属性的嵌入。该算法并未利用属性的具体值，而只利用了属性的名称。

c) 根据属性嵌入计算实体之间的相似度，之后将得到的相似度矩阵与结构嵌入结合得到最终的嵌入，根据最终的嵌入进行对齐。

文献[31]提出了 CTEA(Context and Topic Enhanced Entity Alignment)框架，如图 11 所示。

CTEA 框架包括三个步骤：

a) 根据实体的属性信息为每个源实体生成候选集。具体地，该框架在实体的属性信息中捕获实体主题信息，使用主题模型 BTM(Biterm Topic Model)[92] 的改进版 BTM4EA 对实体主题进行建模，之后根据 JS 距离（Jensen-Shannon divergence）选出候选集。由于属性值的复杂性，该框架没有利用属性的具体值。

b) 根据实体的结构信息生成实体结构嵌入和实体上下文表示。对于实体结构嵌入，该框架使用的是 TransE 模型，根据关系三元组生成实体的结构嵌入，其中已对齐的实体对共享相同的嵌入；对于实体上下文表示，该框架使用的是引入注意力机制的多通道卷积神经网络 CNN(Convolutional Neural Networks)模型，根据实体的邻接信息生成实体的上下文表示，其中模型的输入还包括实体和关系的结构嵌入，用于对模型的参数进行优化。

c) 根据实体结构嵌入和上下文表示分别计算出对应的相似度矩阵，将两个相似度矩阵加权相加得到最终的相似度矩阵，进而在对应的候选集中找出对齐结果。

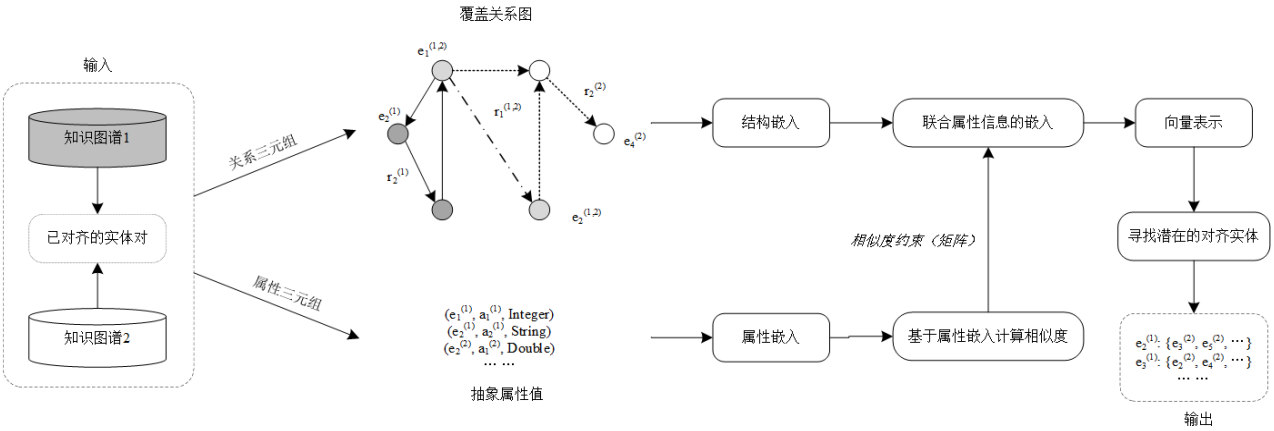


图 10 JAPE 框架图

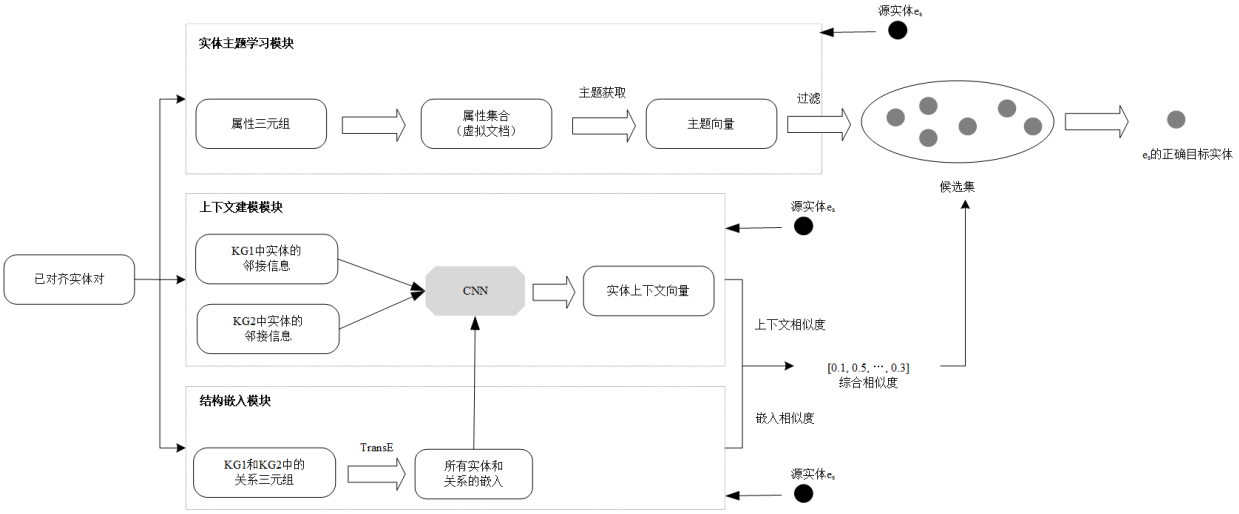


图 11 CTEA 方法框架

(2) 引入属性的具体值

一些方法在利用结构信息和属性名称的基础上，进一步引入了属性的具体值。

文献[24]提出了 SEEA(Self-learning and Embedding based method for Entity Alignment)算法，该算法将实体对齐看做是一种特殊的跨网络关系，并且该算法中仅有这一种关系类型。该算法根据关系三元组和属性三元组对实体、属性和属性值进行嵌入，同时对属性引入了权重机制，重要的属性会拥有更高的权重，之后通过迭代机制进行实体对齐，根据算法对齐的结果对嵌入进行迭代更新。

文献[19]也提出一种 AttrE 方法对属性值进行字符级嵌入，这样在训练时就能够处理没有见过的属性值。AttrE 框架如图 12 所示。

具体地，AttrE 算法包括三个部分：

a) 谓词（关系）对齐：找出两个知识图谱中含义相同的关系，例如“bornIn”和“wasBornIn”，这里使用的是编辑距离且设置相似度阈值为 0.95，之后将含义相同的关系重新命名为相同的名字，从而使得两个知识图谱的关系嵌入在同一个向量空间中。

b) 嵌入学习：对于结构嵌入，该算法使用 TransE 模型，且在学习结构嵌入时更重视包含已对齐谓词的三元组，具体做法是在学习结构嵌入时增加权重机制。对于属性嵌入，该算法把属性三元组也看作是翻译

的过程，h 表示实体，将属性名称看作是谓词（关系）r，由于相同的属性值在不同知识图谱中的表示形式可能不同，例如 50.9989 和 50.9988889，因此该算法对属性值进行字符级的嵌入，之后使用函数 $f_a(\alpha)$将所有的字符嵌入合并为一个向量，相似的属性值具有相似的向量表示，其中 α表示属性值的所有字符嵌入，如下公式(38)：

$$h + r \approx f_a(\alpha) \tag{38}$$

该算法定义了三个函数，第一个是 SUM，即将所有字符嵌入相加；第二个是 LSTM，即使用 LSTM 网络对字符嵌入进行编码；第三个是 N-gram。最后，通过在训练时尽量让同一个实体的结构嵌入和属性值嵌入接近。

c) 实体对齐：使用余弦相似度根据实体的嵌入进行对齐。

此外，该算法还使用了传递性规则进一步丰富实体属性信息，使得属性嵌入更加科学。然而，这种算法在处理跨语言的实体时可能会导致错误。

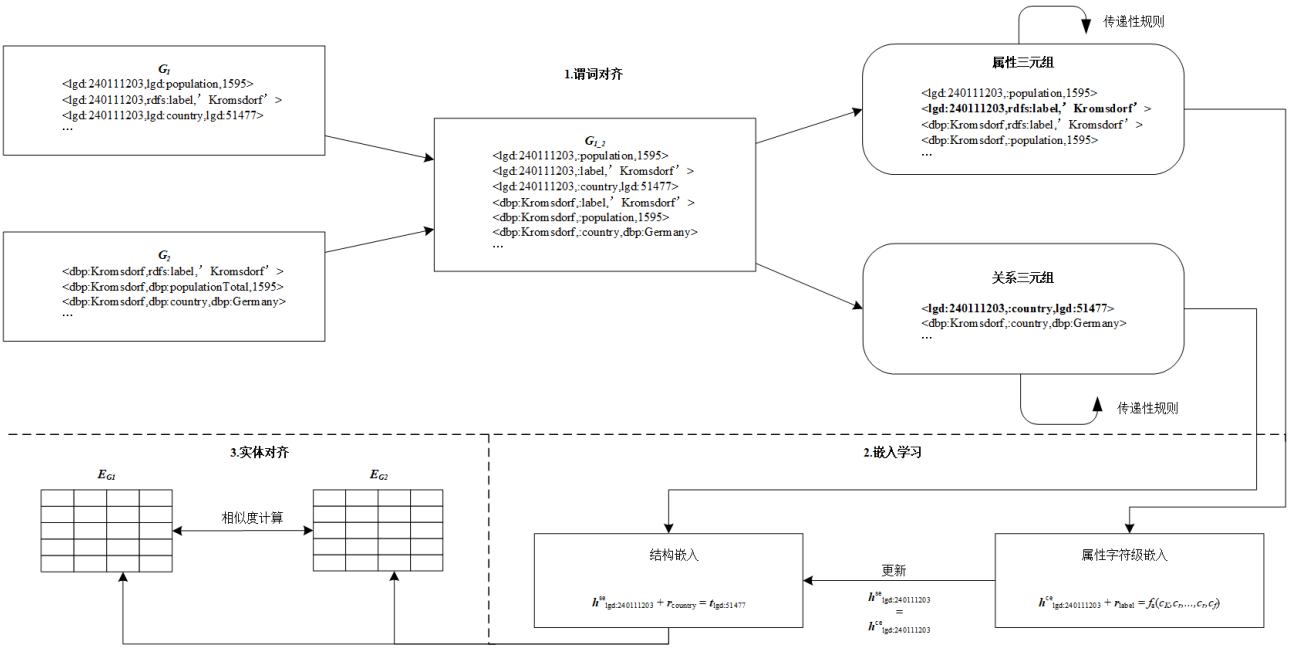


图 12 AttrE 方法框架

文献[4]提出了将实体语义信息同结构信息、属性信息结合的方法。使用协同训练框架，将特征分为语义视角和结构视角，在两个视角下分别训练基于两个图谱联合表示学习(Joint Embedding)的实体对齐模型，不同视角下的实体对齐结果互相辅助参考，实现语义信息和属性信息的融合，从而提升实体对齐的效果。同时，提出使用属性强约束，限制协同训练过程中产生的漂移，即根据表示学习和人工指定的部分属性对实体对进行评分，选出最可能对齐的实体对。该算法除了利用语义信息外，还利用了属性信息，并在属性多的数据集上能够获得更好的效果。之后，该作者在文献[27]中又提出了一个算法，与文献[4]中算法不同的是，该算法提出自适应的决策树模型，根据任务本身的特征选择不同部分信息占据的权重，不再需要人工指定属性进行评分，减少了人工先验知识的不准确性，节约了人力成本。

文献[71]提出使用 GCN 将实体结构和实体属性进行联合嵌入，分别构建了实体的结构特征和属性特征，作为 GCN 的输入。文献[72]提出 COTSAE(CO-Training of Structure and Attribute Embeddings)模型，使用 TransE 进行结构嵌入，将属性类型和属性值进行字符级别嵌入，同时加入联合注意力机制区分重要程度。

文献[73]提出 AttrGNN(Attributed Graph Neural Network)模型，对结构信息和属性信息进行联合编码，使用 GNN 结构嵌入，然后针对实体的不同属性，采用注意力机制获得不同的权重。文献[74]提

出 JarKA(Jointly model the attributes interactions and relationships for cross-lingual Knowledge Alignment)模型，该模型使用一种基于交互的属性模型来捕获属性级别的交互，以估计实体的相似性。

4.2.4 利用结构信息和实体名信息进行实体对齐

目前，大部分实体对齐方法都是根据结构信息和属性信息进行推断，但是在真实世界知识图谱中大部分实体含有的结构信息和属性信息往往很有限。为此，一些方法提出进一步使用实体名信息来进行实体对齐。

文献 **Error! Reference source not found.**提出了 Cross-KG(Cross-Knowledge Graph)算法，该算法利用知识图谱中的结构信息和实体名信息进行对齐。此外，作者指出通过合并稀疏知识图谱和稠密知识图谱的信息，可以实现对稀疏知识图谱进行更科学地嵌入。该算法将稠密知识图谱作为源知识图谱 G_1，将稀疏知识图谱作为目标知识图谱 G_2，Θ 为两个知识图谱中的所有实体和关系的嵌入，该算法定义的目标函数如下公式(39)：

$$\begin{aligned} &P(G_1, G_2|\Theta) \\ &= P(G_1|\Theta)P(G_2|G_1, \Theta) \end{aligned} \tag{39}$$

具体地，该算法包括三个步骤：

(1) 对源知识图谱中的实体和关系进行嵌入。如下公式(40)：

$$P(G_1|\Theta) = \prod_{(h,r,t)\in G_1} P\big((h,r,t)|\Theta\big) \tag{40}$$

其中，(h, r, t)为源知识图谱 G_1 中的关系三元组，分别为头实体、关系和尾实体。在嵌入时，该算法使用的是翻译模型，由于篇幅的限制，该文献只探讨了模型为 TransE 和 TranSparse 时的有效性。

(2) 通过整合源知识图谱的信息，对目标知识图谱中的实体和关系进行嵌入。目标知识图谱中的部分实体与源知识图谱的实体存在链接，记为 M={(e, f)|e 为 G_1 中对应的实体，f 为 G_2 中对应的实体}，如下公式(41)：

$$\begin{aligned} &P(G_2|G_1, \Theta) \\ &\propto \prod_{(s,p,o)\in G_2} P\big((s,p,o)|\Theta\big) \prod_{(e,f)\in M} P(f|e, \Theta) \end{aligned} \tag{41}$$

其中，(s, p, o)为目标知识图谱 G_2 中的关系三元组，分别为头实体、关系和尾实体。该算法只利用了两个知识图谱的实体之间的链接，因为关系之间的链接能够被关系三元组的嵌入捕捉到。

(3) 将实体的结构信息和字符串匹配信息结合起来，计算不同知识图谱的实体之间的相似度。对于字符串匹配信息，该算法使用 Jaro-Winkler 距离(一个度量两个字符序列之间的编辑距离的字符串度量标准)计算实体名字之间的相似度；对于结构信息，该算法定义了两种链接相似度，分别根据连入的关系三元组和连出的关系三元组计算相似度。最后的综合相似度如下公式(42)：

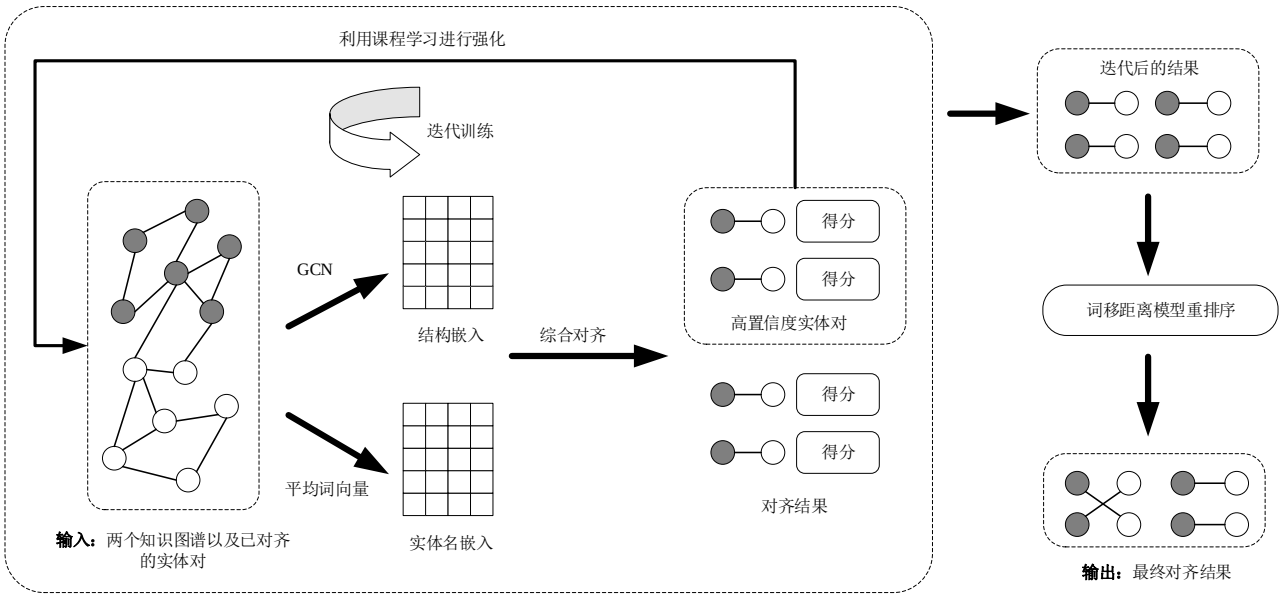
$$\begin{aligned} &SIM(f, e) \\ &= \alpha \\ &* \max\big(sim_{in}(f, e), sim_{out}(f, e)\big) \\ &+ (1-\alpha) sim_{Jaro\_Winkler}(f, e) \end{aligned} \tag{42}$$

其中 f 和 e 分别为两个知识图谱 G_2 和 G_1 中的某个实体，α 为超参数。

文献[22]也引入了实体名信息，将其与结构信息结合进行实体对齐，提出了一种基于重排序的迭代式实体对齐方法，如图 13 所示。具体地，首先，利用图卷积网络 GCN(Graph Convolutional Network)捕捉结构特征得到结构嵌入，再利用平均词向量的方法捕捉实体名特征得到实体名嵌入，其中词向量模型使用的是 FastText[93]。之后，根据结构嵌入和实体名嵌入分别计算出结构嵌入的距离和实体名嵌入的距离，将这两个距离加权相加得到实体间的距离，进而实现实体对齐。接着又提出了基于课程学习的迭代策略，模仿人类学习的特点，从易至难地选择高置信度实体对集合对训练集进行扩充，其中的难易程度由实体节点度数高低来刻画，度数较高的实体的结构信息更丰富，更易于对齐，而度数较低的实体则相对而言有一定难度。最后，在迭代结束后，使用词移距离模型进一步对实体名信息进行挖掘，将对齐结果进行重排序，提升实体对齐的准确率。具体地，在迭代结束时，为每一个待对齐实体保留多个候选实体，之后利用词移距离模型重新计算出实体名之间的距离，再结合结构嵌入计算出最终的对齐结果。

文献 [81] 提出 DAT(Degree-Aware Entity Alignment In Tail)模型，使用了 RSNs(Recurrent Skipping Networks)[58]知识表示模型进行结构嵌入，将实体名称进行了字符级别嵌入，通过计算两个实体的结构相似性和实体名称相似性，并采用注意力机制将两者按照权重结合。文献 [82] 提出 CEA(Collective Entity Alignment)模型，使用 GCN 进行结构信息嵌入，并将实体名称词向量取平均值进行嵌入。

图 13 基于重排序的迭代式实体对齐方法框架



4.2.5 利用结构信息和实体描述信息进行实体对齐

一些实体对齐方法也充分利用了实体的描述信息。部分知识图谱提供了**实体的描述信息**，例如 Wikidata 中实体“bus”的描述信息为“large road vehicle for transporting people”。

文献[28]提出了一种结合结构信息和实体描述信息进行跨语言实体对齐的模型。主要分为三部分。

(1) 该模型利用 TransE 训练得到基于结构信息的知识向量，之后利用共享参数模型将不同知识图谱的知识向量集成到同一个语义空间，共享参数模型的策略是使得不同知识图谱的已对齐实体共享相同的嵌入，进而找到可能被对齐的实体对。

(2) 结合实体描述信息选出最终的对齐实体。具体地，首先使用 CBOW(Continuous Bag Of Words)模型分别训练各语言的词向量，再学习不同语言词向量之间的线性映射，将不同语言的词向量空间集成到同一个语义空间中，之后使用改进后的最优对齐方法来计算不同语言实体描述之间的相似度，进而选出最终的对齐实体。

(3) 通过迭代对齐的方式重复前两个步骤，以找到更多的实体对。其中，为了缓解错误传播问题，该模型采用了两种策略，一是每一轮迭代中都将知

识向量重新初始化再进行训练，二是软对齐，即为每一个新对齐的实体对设置一个评分。

4.2.6 利用结构信息、属性信息和实体描述信息进行实体对齐

文献[83]提出使用多方面信息综合进行实体对齐，如：实体标签、拓扑信息、关系类型、属性和文本描述等。在利用 GCN 嵌入结构信息的同时加入多方面信息，提出 HMAN(Hybrid Multi-Aspect Alignment Network)模型，将拓扑信息、关系和属性这三方面信息同时作为 GCN 的输入，而不是仅仅作为外部辅助信息。最后将文本描述信息利用 BERT 嵌入学习，将结构嵌入和文本描述信息嵌入相结合，综合进行实体对齐，模型如图 14 所示。使用多层 GCN 来获取知识图谱的拓扑信息，使用全连接层和高速门机制来进行关系和属性嵌入，最后将拓扑信息、关系信息和属性信息连接起来。

4.2.7 利用结构信息、属性信息和实体名信息进行实体对齐

文献[23]提出了 MultiKE(Multi-view KG Embedding)框架，同时利用了结构信息、属性信息和实体名信息。对于实体名，首先对其中的每个单词分别进行嵌入，为了捕捉单词在不同表达方式下的语义信息，如果单词有词嵌入则使用词嵌入，否则使用字符嵌入的平均作为嵌入，最后将所有单词的嵌入进行连接再以无监督的方式压缩为一个名称嵌入。对于实体之间的关系，该框架使用 TransE 得到关系嵌入。对于实体的属性，该框架使用卷积神经网络 CNN(Convolutional Neural Networks)捕捉属性及其值中的信息，首先将属性名嵌入和属性值嵌入横向连接起来，再输入到 CNN 中得到压缩后的属性嵌入。

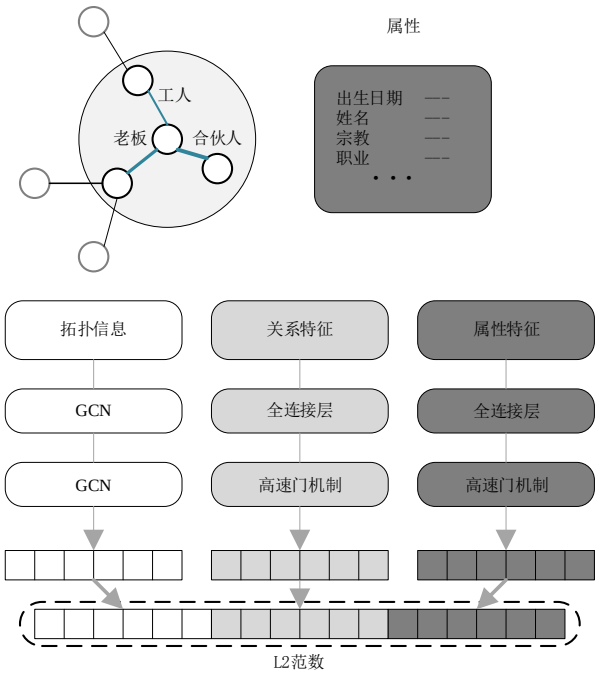


图 14 HMAN 模型图

为了将这三个嵌入组合生成实体的嵌入，该框架提出了三种策略。第一种是基于权重机制进行组合，将三个嵌入加权相加作为实体的嵌入。第二种是基于共享空间进行组合，将每个嵌入的空间通过正交映射矩阵映射到同一个空间中。第三种则是在训练时对所有嵌入进行联合训练，使得在同一个嵌入空间中的实体嵌入和这三个嵌入之间的一致性尽可能大。

此外，为了更好地学习两个知识图谱之间的联系，该框架提出了两种跨知识图谱的推理方法，根据已对齐的实体、关系和属性进行推理，以生成更多的三元组。

文献[84]提出 EPEA(entity-pair embedding approach)模型，用成对实体连接图(pair-wise connectivity graph, PCG)来学习实体嵌入表示，使用卷积神经网络(CNN)进行属性特征提取，进一步构建边缘注意力的 GNN 传递实体间特征相似性。

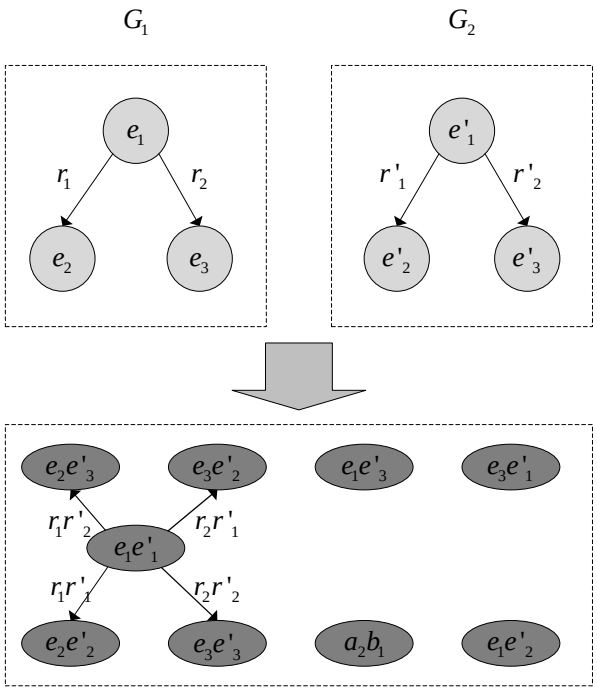


图 15 成对连接图(PCG)

如图 15 所示，表示两个知识图谱 G_1 和 G_2 的成对连接图 PCG，e_i和r_i表示 G_1 中的实体和关系，e'_i和r'_i表示 G_2 中的实体和关系。PCG 的节点是由两个知识图谱中的实体对组成，而边则是由关系对组成。在实际操作中，为避免产生不必要的实体对节点，EPEA 模型只考虑实体名和属性特征高度相似的实体对。

在 PCG 生成实体对之后，通过 CNN 自动捕获实体之间的属性特征相似性，将属性特征矩阵作为 CNN 的输入，学习实体对的嵌入向量表示。EPEA 针对两个实体的属性对统一看作字符串处理，计算属性值的 Jaccard 相似性。而实体名被作为一种特殊的属性，同 CNN 的输出特征进行连接。

最后，构建基于注意力机制的特征传播。利用 GNN 加入边缘注意力机制，通过递归聚合邻居节点的特征向量学习节点的向量表示，这些特征向量能够将图中的节点特征和结构信息结合起来。为了进一步强化实体嵌入时原始属性特征，在 GNN 模型的输出层加入了剩余连接。

如图 16 所示是特征提取和特征传播的框架图，A_i代表 G_1 中e_i实体的属性值，A'_i代表 G_2 中e'_i实体的属性值，x_i表示 CNN 输出特征向量，z_i表示实体名特征，h_i表示 GNN 的节点特征，h'_i经过 GNN 之后生成的新的节点特征。

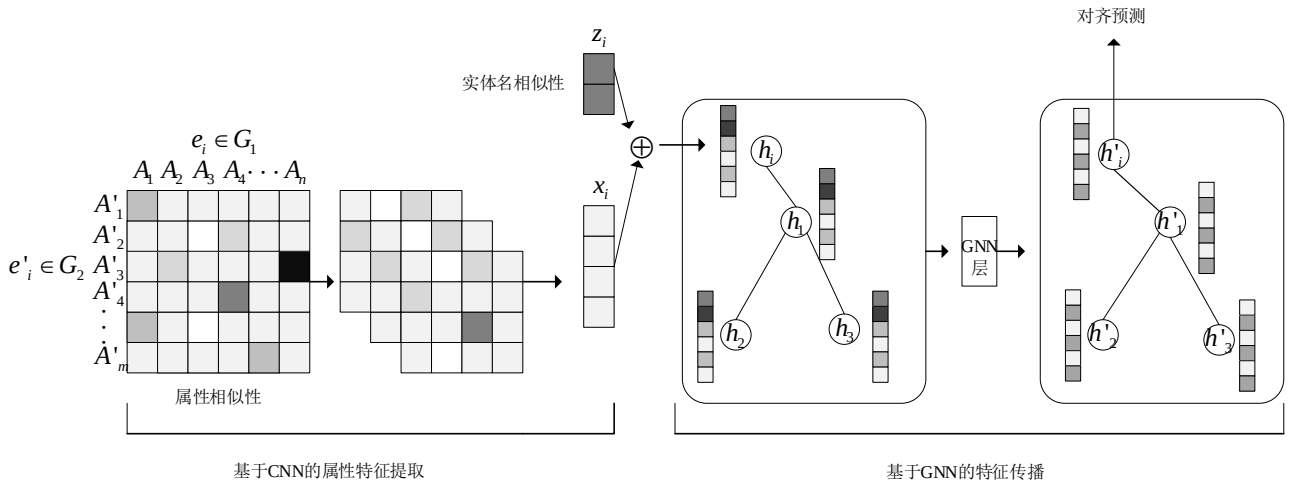


图 16 特征提取和特征传播框架图

4.2.8 利用结构信息、属性信息、实体名信息和实体描述信息进行实体对齐

文献 [85] 综合以上所有信息，提出 BERT-INT(BERT-based INTeraction model)，与文献[70]不同的是，并没有使用属性信息作为 BERT 的输入，而是使用实体名称和实体描述信息作为输入训练实体的嵌入表示，利用实体之间的邻接信息、关系和属性信息计算相似性。与文献[83]的不同在于没有将所有信息来训练嵌入表示。主要分为两个模块：BERT 嵌入模块（BERT-INT-1，如图 17 所示使用实体名称和实体描述信息作为输入信息训练

嵌入向量）和基于 BERT 的交互模块（BERT-INT-2，如图 18 所示）。

BERT-INT 模型首先使用 BERT 作为基本的表示单元来嵌入实体的名称/描述、属性和值，并建立邻居视图和属性视图交互模型来计算这些嵌入之间的交互。假设两个实体集E和E'分别来自不同的知识图谱，其中$\{r_1,\dots,r_n\}$和$\{e_1,\dots,e_n\}$表示实体集E的关系和实体，$\{r_1',\dots,r_n'\}$和$\{e_{1'},\dots,e_n'\}$表示实体集E'的关系和实体；$\{a_1,\dots,a_n\}$和$\{v_1,\dots,v_n\}$表示实体集E的属性和属性值，$\{a_1',\dots,a_n'\}$和$\{v_1',\dots,v_n'\}$表示实体集E'的属性和属性值。在属性信息交互模块，需要输入候选实体对的属性和属性值，$C(a_i)$和$C(v_i)$；在邻接信息交互模块需要输入关系和实体，$C(r_i)$和$C(e_i)$。最终基于 BERT 的交互模块结合了实体邻接信息、邻接实体关系和属性信息综合计算实体相似性来完成实体对齐。

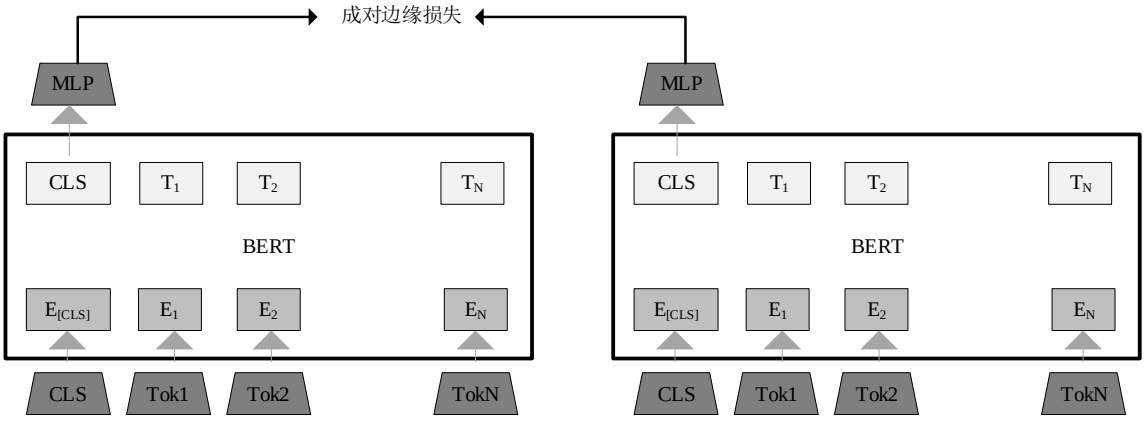


图 17 BERT-INT-1 模型图

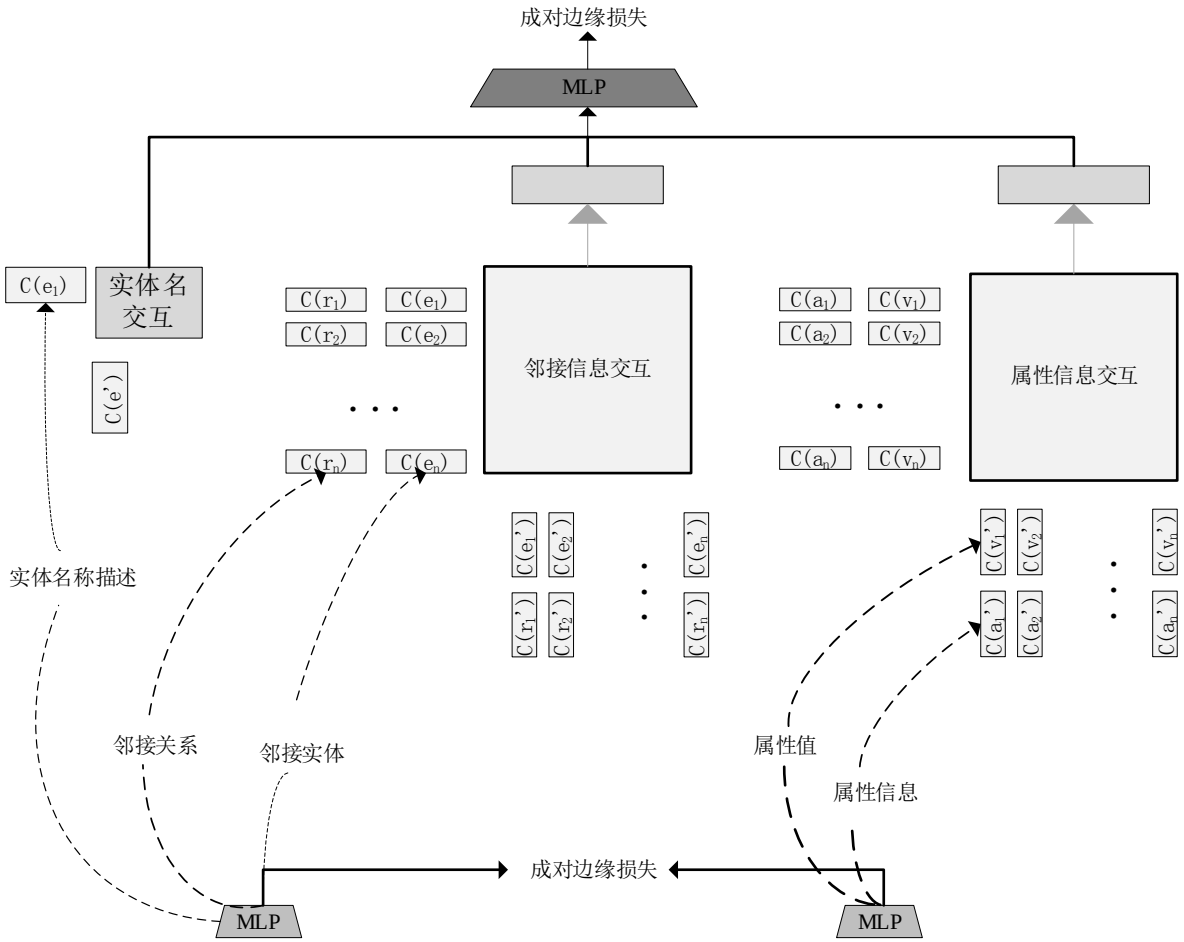


图 18 BERT-INT-2 模型图

4.3对比分析

本节针对基于通用数据集(DBP15K)和非通用数据集的实体对齐方法进行总结(如表 8 和表 9 所示)，并进一步对基于通用数据集的不同方法模型的实验效果和模型效率进行深入对比分析。

4.3.1 模型效果对比

首先，根据所利用信息的不同，将基于知识表示学习的实体对齐方法分为只利用结构信息、利用结构信息和额外信息(属性信息、实体名信息、实体描述信息等)、只利用额外信息三类。然后，按照嵌入方法的不同，进一步细分为基于 TransE、GCN 和 GNN 三类。

只利用结构信息：

(1) 基于 TransE 结构信息的利用，TransEdge(2020)[66]获得了最好的效果，说明对知识图谱的嵌入表示时，不仅仅要对头尾实体进行向量嵌入，加入对实体之间的关系嵌入可以更好地表示知识图谱。TransEdge(2020)[66]、Multi-mappingRelations(2019)[62] 和 KECG(2019)[64]均对实体之间的关系进行了考虑，三者之中利用 GAT 标注关系权重的 KECG(2019)[64] 获得的实验效果最差。而 Multi-mappingRelations(2019)[62]考虑了实体之间的多关系，在 MTransE(2017)[60]的基础上通过对 TransE 得分函数和损失函数的改进，同时在嵌入表示时加入 ComplEx 对多关系进行语义表示，从而在实验效果上获得了显著提高。BootEA(2018)[61]通过加入自举训练的策略，来缓解训练过程中标注数据缺乏的问题，虽然缺

乏对实体之间的关系考虑，但是同KECG(2019)[64]相比获得了更好的效果。

(2) 基于GCN结构信息的利用，SSP(2020)[69]的实验表现效果明显优于AliNet(2020)[70]。SSP(2020)[69]不仅对全局结构信息进行考虑，同时考虑实体之间的关系，结合翻译模型加入更细粒度的关系语义信息，而AliNet(2020)[70]并没有加入关系信息，而是对多跳实体区分考虑，但对直接邻接实体则统一对待，表明在结构信息的利用中，加入对实体之间关系的考虑在实体对齐任务中可以获得更好的效果。

(3) 基于GNN结构信息的利用，MuGNN(2019)[63]，MRAEA(2020)[68]和RREA(2020)[79]均对实体之间的关系进行了考虑，其中RREA(2020)[79]的实验效果最好。MuGNN(2019)[63]和MRAEA(2020)[68]均使用注意力机制对实体之间的关系分配权重，MRAEA(2020)[68]针对实体之间的关系进一步对类型、方向和可逆性进行了考虑，并加入多头注意力机制，所以在实验效果上优于MuGNN(2019)[63]。此外MuGNN(2019)[63]在训练中不仅仅需要预先对齐实体，还需要对齐关系，在标记训练数据时耗费时间和人力。RREA(2020)[79]通过对多跳邻接实体的考虑，进一步对实体之间的关系转换映射，仅利用结构信息就获得了较好的对齐效果。

从以上分析可以看出，仅利用结构信息的实体对齐方法，基于TransE的嵌入方法整体上效果表现较差，主要原因是使用TransE构建关系三元组，缺乏对知识图谱整体结构信息的考虑，从而丢失部分结构信息。此外，从基于GCN/GNN的方法可以看出，相比于MuGNN(2019)[63]和MRAEA(2020)[68]，RREA(2020)[79]在利用结构信息的实验效果上获得了最好的结果，说明基于GCN或者GNN的嵌入表示方法，在表示实体和关系嵌入时，若不加任何限制和约束，会导致变换后的结果破坏原有知识图谱结构信息的相似性，从而在实体对齐任务上表现较差。

利用结构信息和额外信息：

(1) 基于TransE的嵌入方法中，CTEA(2020)[31]在Hits@10上表现最好，表明使用卷积神经网络学习邻接实体的上下文信息，进而计算实体相似性发挥了一定的作用。JarKA(2020)[74]对于属性信息的利用上不同于JAPE(2017)[18]，后者在嵌入表示时加入属性三元组联合学习表示向量，而JarKA(2020)[74]则利用属性三元组以及属性值信息训练机器翻译模型获得高置信度的预对齐实体，加入结构嵌入训练中，从而获得了更好的实验表现效果。

(2) 基于GCN的嵌入方法中，GCN(SE+AE)(2018)[71]使用GCN对结构信息和属性信息进行嵌入，缺乏对实体之间的关系考虑，将所有邻接实体统一对待，因此实验效果最差。其中RDGCN(2019)[65]、HGCN(2019)[63]、NMN(2020)[72]和CEA(2020)[82]使用的额外信息均为实体名，并且都在初始化时采用了实体名的预训练词向量。RDGCN(2019)[65]和NMN(2020)[72]都采用了注意力机制，由于NMN(2020)[72]通过划分子图采用跨图注意力机制，选择性考虑中心实体的邻接实体，而RDGCN(2019)[65]对所有邻接实体之间的关系分配权重，导致了错误传播，因此表现效果较差。NMN(2020)[72]不仅仅在初始化时使用实体名信息，并且使用实体名预训练向量对邻接实体进行筛选，实验效果优于RDGCN(2019)[65]和HGCN(2019)[63]，表明并不是所有直接相邻实体都对实体嵌入表示有贡献。而CEA(2020)[82]提出在使用实体名嵌入的同时，加入了实体名字符串处理，并且提出了延迟接受算法(deferred acceptance algorithm, DAA)来进行稳定匹配，取得了更好的效果。尤其在DBP15KJA-EN的表现上得到了很大提高，这表明基于实体名字符串的方法，在语言差异小的数据集上更加适用。RDGCN(2019)[65]、HGCN(2019)[63]、RNM(2021)[80]和HMAN+BERT(2019)[87]均对实体之间的关系进行表示嵌入，并且都使用了高速门机制缓解GCN中的错误传播，而HGCN(2019)[63]和RNM(2021)采用了自举训练策略，因此HGCN(2019)[63]的效果略优于RDGCN(2019)[65]，而RNM(2021)[80]不仅仅使用关系信息进行嵌入，并且通过划分子图，同时匹配邻接实体和关系，放大关系信息在结构中的作用，在实验效果上获得了显著提高。HMAN+BERT(2019)[83]不仅仅考虑了实体之间关系，实体属性信息，还加入了实体的描述信

息，并利用 BERT 预训练模型训练嵌入向量，在实验效果上获得了显著提高，这得益于额外信息和 BERT 的利用。

(3) 基于 GNN 的嵌入方法中，EPEA(2020)[84]的表现效果最好，这表明成对连接图(PCG)、CNN 捕获属性特征，再加以 GNN 边缘注意力传播，这种嵌入方法递归聚合邻居节点的特征向量，可以兼顾属性信息和结构信息，很好的利用了知识图谱中的数据信息。而加入了实体名信息的 RREA(text)(2020)[79]表现次之，表明使用 GNN 进行嵌入变换操作的约束限制，可以很好的保留原有知识图谱的结构信息。DGMC(2020)[77]利用软对齐和同步消息传递策略，在 Hits@1 的实验效果略优于 AttrGNN(2020)[73]，而在 Hits@10 上较差。表明 AttrGNN(2020)[73]通过属性信息划分子图，然后根据子图来嵌入的方法，可以将实体对齐的标准结果很好的划分在一定区域，但是准确性不够，因此 Hits@1 的得分较低。

基于以上分析，利用结构信息和额外信息的实体对齐方法，基于 GNN 的嵌入方法加入实体额外信息之后的整体表现较好。在考虑结构信息和属性信息来看，GCN 的表现效果略优于 TransE，其中 HMAN+BERT(2019)[83]的实验效果表现最好，分析可知，只有 HMAN+BERT(2019)[83]使用了实体的描述信息，并采用 BERT 对描述信息进行训练，而实体的描述信息来自数据库 DBpedia，通过描述信息可以很好的获取实体的语义信息，从而获得更好的区分度。

只利用额外信息：

从表 8 中可以看到，BERT-INT(2020)[85]在整体效果上获得了显著提高，表明通过实验数据集之外的额外信息，能够很准确的获得实体语义信息。HMAN+BERT(2019)[83]在使用结构信息的基础上，也使用了额外描述信息，但效果远不如 BERT-INT(2020)[85]，这表明额外信息比结构信息更利于实体对齐。

此外，由表 8 可以看到，对于使用预训练词向量的模型，如 RDGCN(2019)[65]、HGCN(2019)[75]、NMN(2020)[72]、RNM(2021)[80]、HMAN+BERT(2019)[87]、DGMC(2020)[74]、AttrGNN(2020)[82]、RREA(text)(2020)[75]和 BERT-INT(2020)[88]，这些模型的实验效果在 DBP15KFR-EN 的效果明显高于 DBP15KZH-EN 和 DBP15KJA-EN，这表明法语和英语的相似性，使得预训练词向量在结果上更加接近，从而匹配效果较好。

表 8 基于 DBP15K 数据集的实体对齐方法

嵌入方法	模型	$DBP15K_{ZH-EN}$		$DBP15K_{JA-EN}$		$DBP15K_{FR-EN}$	
		Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
只利用结构信息							
TransE	MTransE(2017)[60]	30.83	61.41	27.86	57.45	24.41	55.55
	KECG(2019)[64]	47.77	83.50	48.97	84.40	48.64	85.06
	BootEA(2018)[61]	62.94	84.75	62.23	85.39	65.30	87.44
	Multi-mappingRelations(2019)[62]	68.07	86.74	65.53	85.90	67.70	80.84
	TransEdge(2020)[66]	73.5	91.9	71.9	93.2	71.0	94.1
GCN	AliNet(2020)[65]	53.9	82.6	54.9	83.1	55.2	85.2
	SSP(2020)[65]	73.9	92.5	72.1	93.5	73.9	94.7
GNN	MuGNN(2019)[63]	49.4	84.4	50.1	85.7	49.5	87.0
	MRAEA(2020)[68]	75.70	92.98	75.78	93.38	78.04	94.81
	RREA(2020)[79]	80.1	94.8	80.2	95.2	82.7	96.6
利用结构信息和额外信息							
TransE	JAPE(2017)[18]	41.18	74.46	36.25	68.50	32.39	66.68
	JarKA(2020)[74]	70.58	87.81	64.58	85.50	70.41	88.81
	CTEA(2020)[31]	-	90.5	-	91.4	-	92.3

GCN	GCN(SE+AE)(2018)[71]	41.25	74.38	39.91	74.46	37.29	74.49
	RDGCN(2019)[76]	70.75	84.55	76.74	89.54	88.64	95.72
	HGCN(2019)[75]	72.03	85.70	76.62	89.73	89.16	96.11
	NMN(2020)[78]	73.3	86.9	78.5	91.2	90.2	96.7
	CEA(2020)[82]	78.7	-	86.3	-	97.2	-
	RNM(2021)[80]	84.0	91.9	87.2	94.4	93.8	98.1
	HMAN+BERT(2019)[83]	87.1	98.7	93.5	99.4	97.3	99.8
GNN	AttrGNN(2020)[73]	79.60	92.93	78.33	92.08	91.85	97.77
	DGMC(2020)[77]	80.12	87.47	84.80	89.74	93.34	96.03
	RREA(text)(2020)[79]	82.2	-	91.8	-	96.3	-
	EPEA(2020)[84]	88.5	95.3	92.4	96.9	95.5	98.6
只利用额外信息							
BERT	BERT-INT(2020)[85]	96.8	99.0	96.4	99.1	99.2	99.8

表 9 非通用数据集的实体对齐方法

模型	嵌入方法	数据集来源	评价指标
MTransE (2017)[60]	TransE	DBpedia、WK3l	Hits@10、MR
SEEA(2019)[24]	TransE	Cora、百度、豆瓣	Precision、Recall、F1-measure
AttrE(2019)[19]	TransE	DBpedia、LinkedGeoData Geonames、YAGO	Hits@k、MR
融合语义和结构信息的实体对齐方法(2019)[4]	TransE	Cora、百度、豆瓣	Precision、Recall、F1-measure
AKE(2019)[86]	TransE	DBpedia	Hits@k、MR
VR-GCN(2019)[30]	GCN	DBpedia	Hits@k、MRR
REA(2020)[91]	GNN	DBpedia	Hits@k、MRR
自适应属性选择的实体对齐方法(2020)[94]	TransE	Cora、百度、豆瓣	Precision、Recall、F1-measure
Cross-KG(2017)Error! Reference source not found.	TransE	DBpedia	Hits@ k
基于重排序的迭代式实体对齐(2020)[22]	GCN	Cora、百度、豆瓣	Precision、Recall、F1-measure
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(2019)[28]	TransE	DBpedia	Hits@k、MR
MultiKE(2019)[23]	TransE	DBpedia、Wikidata、YAGO3	Hits@k、MR、MRR
OTEA 最优运输 2019[88]	TransE	DBpedia、WK3l	Hits@k、MRR
SEA(2019)[90]	TransE	DBpedia、WK3l	Hits@k、MRR
CG-MuAlign(2020)[26]	GNN	IMDB、Freebase	Hits@1、PRAUC

		Wikipedia Amazon Music	(Precision-Recall Area Under Curve) F1-measure
SelfAttention-GCN(2020)[69]	GCN	DBPedia、YAGO3、Wikidata	Hits@k、MRR
Schema-Agnostic(2020)[70]	BERT	the publicly available datasets on Github[10]	Precision、Recall
COTSAE(2020)[72]	TransE	DBpedia,、Wikidata、YAGO3	Hits@k、MRR
DAT(2020)[81]	RSNs	DBpedia	Hits@k、MRR

[10]https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

此外，基于注意力机制在知识图谱各项任务中的重要作用，本文对基于知识表示学习的实体对齐方法中采用了注意力机制的方法也进行了对比(如表 10 所示)。需要说明的是，这些方法并未在相同数据集上进行实验，因此目前表 10 仅列出了每种方法在使用注意力机制时的异同，并未给出实验性能数据。在未来工作中，本文将进一步进行基于注意力机制的实体对齐实验对比分析。

表 10 采用注意力机制的实体对齐方法

模型	嵌入方法	注意力权重
RDGCN(2019)[76]	GCN	图注意机制(GAT)，给邻接实体分配权重
MuGNN(2019)[63]	GNN	自注意力机制和知识图谱交叉注意力，给邻接实体和关系分配权重
KECG(2019)[64]	TransE	自注意力机制，给相关实体分配权重
CTEA(2020)[31]	TransE	在卷积层加入注意力机制给邻接实体分配权重
AliNet(2020)[65]	GCN	注意力机制，给远距离实体分配权重
MRAEA(2020)[68]	GNN	自注意力机制，给关系分配权重
NMN(2020)[78]	GCN	跨图注意力机制，给邻接实体分配权重
CG-MuAlig(2020)[26]	GNN	节点级别注意力机制和边级别注意力机制
SelfAttention-GCN(2020)[69]	GCN	多头注意力机制，给长距离实体分配权重
COTSAE(2020)[72]	TransE	双层 GRU，根据属性类型和属性值使用联合注意力，给属性分配权重
AttrGNN(2020)[73]	GNN	注意力机制，给属性信息分配权重
DAT(2020)[81]	RSNs	注意力机制，给结构相似性和实体名相似性分配权重
EPEA(2020)[84]	GNN	边缘注意力机制，对两实体边缘类型分配权重

4.3.2 模型效率对比

为了对模型的效率进行比较分析，本文将基于通用数据集DBP15K的模型在同一平台上进行了实验。每个模型在 $DBP15K_{ZH\text{-}EN}$ 数据集上运行五次获得平均运行时间，通过运行结果对比分析了不同模型的效率。实验统一使用 IntelXeon E5 2.10GHz 的 CPU、64G 内存、以及一个 NVIDIA GeForce GTX 1080Ti GPU，并在 MANJARO 环境下运行。

如图 19 所示为不同模型在同一数据集 $DBP15K_{ZH\text{-}EN}$ 上运行时间对比。其中蓝色表示只利用结构信息的模型，橙色表示利用结构信息和额外信息的模型，由浅至深分别代表基于 TransE、GCN 和 GNN 的模型。*HGCN、*RDGCN、**RDGCN 分别表示词向量嵌入维度为 100 时的 HGCN、词向量嵌入维度为 100 时的 RDGCN、以及词向量嵌入维度为 200 时的 RDGCN(源代码的嵌入维度均为 300)。

其中 HGCN、RDGCN 和 NMN 的源代码，在运行时出现内存溢出的情况，说明这三个模型的空间复杂度高，对空间的需求大于其他模型。将维度降低为 200 时，只有 RDGCN 可以运行；降低为 100 时 HGCN 可以运行，NMN 仍不可运行。由此也可以间接说明空间复杂度 NMN>HGCN>RDGCN>其他模型。上述结果也与每个模型的特点相吻合，正如 4.3.1 节也提到，其中 NMN 需要划分子图并且对邻接实体采用注意力机制进行筛选，进而需要更大的空间；而 HGCN 采用自举策略进行迭代，扩充种子实体对，因此需要的空间大于 RDGCN。

此外，由图 19 也可以看到，利用结构信息和额外信息的实体对齐模型的运行时间整体上大于只利用结构信息的模型。这表明引入额外信息的同时会使得模型的复杂度变高，从而增加运行时间。

另外，按照知识表示的嵌入方法分类来看，基于 GCN 和 GNN 模型运行时间，要小于基于 TransE 的模型。实验中本文发现基于 TransE 的模型，使用

GPU 资源都较少，较多的运算使用 CPU，速度慢，但对空间需求小；而 GCN 和 GNN 大量使用 GPU 计算，运算速度快，但对空间需求大。

再者，从图 19 中可以看出，使用迭代自举策略的模型运行时间较长，如 BootEA 和 HGCN，这表明迭代机制增加了种子实体对，同时也增加了模型运算量，从而增加搜索复杂度。HGCN 和 RDGCN 的空间复杂度和时间复杂度较高，在降低嵌入维度的情况下运行时间仍然较长，该类模型不适合大型知识图谱的应用场景。

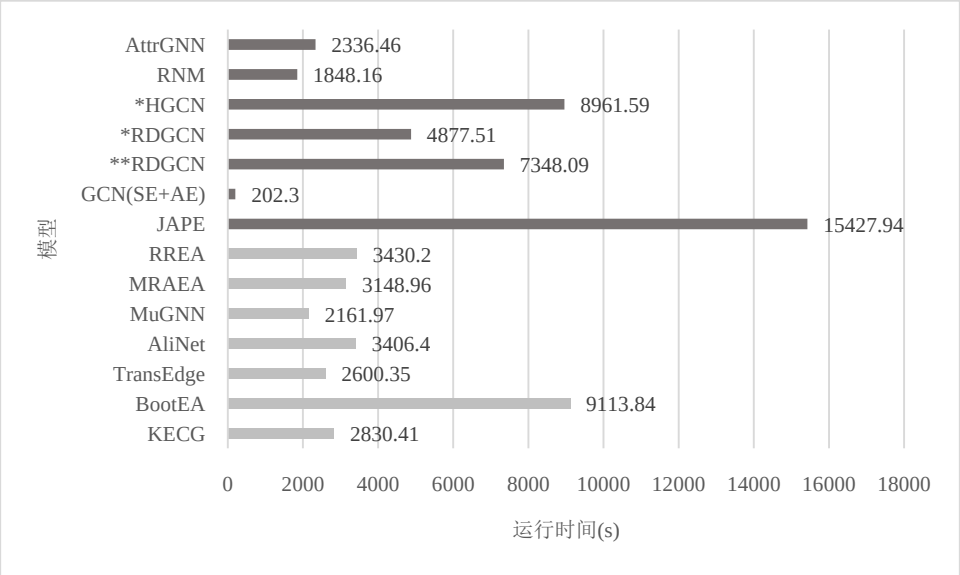


图 19 不同模型在同一数据集 $DBP15K_{ZH\text{-}EN}$ 上运行时间对比

总体来看，综合模型整体效果和时空复杂度来看，RNM 的表现较好。在使用较少空间和较少时间的情况下达到了不错的效果。这表明基于关系的子图匹配，并且对邻接实体进行概率计算，很适合知识图谱实体对齐任务。对于基于 BERT 的两个模型，HMAN+BERT 和 BERT-INT 并未列入图 19 中。因为利用 BERT 加入描述信息进行微调的时间远大于其他模型，使用 BERT 的模型虽然有效地提高了实体对齐效果，但是模型的普适性有待提高，且需要较准确的实体描述信息，这在实际应用中往往不易满足。

5 存在的问题与展望

正如上文所述，近年来实体对齐任务受到了广泛关注，同时也出现了多种实体对齐技术。然而，随着不同领域的应用需求，实体对齐技术仍然存在许多问题亟待解决。

5.1 存在的问题

5.1.1 稀疏知识图谱的处理

目前，主流的实体对齐方法都主要借助知识图谱的结构信息进行对齐，这些方法在人工构建的数据集上取得了最好的实验效果。但是，正如文献[21]中也指出，人工构建的数据集中的知识图谱比真实世界中的知识图谱更稠密。具体来说，在如今存在的知识图谱中，绝大多数实体的邻接实体只有一个或两个，称为长尾实体(long-tail entities)，目前知识图谱中大部分实体都是长尾实体，这是研究实体对齐工作的一大障碍，因为大多模型借助的结构信息不再丰富。

除了借助结构信息，一些方法也引入了属性信息以补充结构信息，这类模型都有一个统一的前提就是属性三元组均存在。但是，正如文献[95]中指出，大多数的知识图谱中实体包含的属性信息存在不同与缺乏的问题。此外，虽然实体描述也能够提供额外信息，但是大多数的实体并没有丰富的描述信息，这就使得这类模型方法有很差的通用性。文献[22]提出了利用实体名称的预训练向量，使得长

尾实体带来的问题得到了一定程度的缓解。此外，在互联网中提取额外的信息以对长尾实体的相关信息进行补充也有可能成为一种缓解方案[94]。但是，到目前为止，还未发现直接解决长尾实体对齐问题的有效方法。

5.1.2 标注数据的缺乏

为了搭建连接两个知识图谱的桥梁，需要大量的标记的预对齐实体对。然而，存在的标注好的数据是少量的，人工标注需要的工作量又往往很大。

为了解决这个问题，一些方法引入了迭代机制，从算法得到的结果中选出高置信度实体对，用于下一轮训练。文献[29]在算法对齐的结果中选出高置信度实体对，不断扩充预对齐实体对。并提出使用软对齐策略缓解错误传播的问题，这使得模型训练时间长，复杂度高。文献[94]提出了自举训练(bootstrapping)框架，为了提高扩充训练集的准确率，该框架使用了全局优化目标，但是这个策略会大大降低算法的效率。

此外，在面向大型真实知识图谱时，这些采用迭代的方法只能得到少量高置信度的实体对，这就导致模型的表现得不到显著地提升。目前，还没有发现直接解决标注数据缺乏问题的方法。

5.1.3 标注数据中的噪声处理

目前的大多数实体对齐方法都需要根据标注数据(已对齐实体对)进行对齐，这些方法都假设标注数据是正确无误的。然而，如图 20 所示，标注数据中也可能会存在错误的数据，即噪声，这在很大程度上影响了对齐的效果。这些噪声数据的来源主要有两个，一个是人工标注时产生的错误数据，另一个是在迭代过程中产生的错误数据，部分实体对齐方法引入了迭代机制以解决标注数据缺乏的问题，但是迭代过程中生成的实体对并不一定都是正确的。

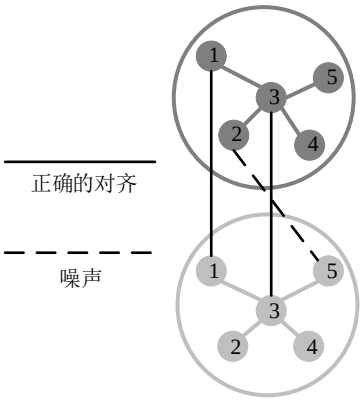


图 20 噪声数据（编号相同的实体具有相同的语义，应该被对齐）

针对噪声数据问题，文献[91]提出了 REA 框架，由于标注数据可能存在错误，该框架为每个实体对都设置一个置信度，该框架包括两个模块：噪声感知实体对齐模块和噪声检测模块。该框架引入了迭代机制，在迭代之前，首先初始化已标注实体对的置信度，具体地，将一定正确的实体对的置信度初始化为 1，将不确定是否正确的实体对的置信度初始化为 0。每轮迭代的步骤如下：

(1) 噪声感知实体对齐模块：使用图神经网络 GNN(Graph Neural Networks)结合实体对的置信度对知识图谱的结构信息进行嵌入，得到知识图谱中实体的嵌入。

(2) 噪声检测模块：该模块首先根据正确的实体对生成一些噪声实体对，之后使用这两部分实体对对噪声检测模型进行训练。这两个步骤是相互补充的，生成的噪声实体对用于噪声检测模型的训练，而噪声检测模型又能对生成噪声实体对的策略进行优化，使得生成的噪声实体对不容易被现有的噪声检测模型识别。

(3) 每轮迭代的最后，使用训练好的噪声检测模型对实体对的置信度进行更新。

进行多轮迭代直到整个模型收敛，最后根据实体的嵌入进行对齐。

该框架能在一定程度上识别出标注数据中的噪声，然而该框架仍然存在着两点不足。第一，在迭代之前需要初始化已标注实体对的置信度，为了保证对齐的效果，此处的已标注实体对只能通过人工的方式标注，将正确的实体对标注出来；第二，噪声检测模型识别的结果并不一定全部正确，从而影响了对齐的效果。

5.2 展望

从上述已有方法可以看出，目前基于知识表示学习的实体对齐方法主要包括三个模块，分别是嵌入模块、交互模块和对齐模块。嵌入模块目前主要有三种方法，一种是利用 TransE 及其改进系列进行关系结构信息嵌入；一种是使用 GNN 构建邻接关系图进行嵌入；一种是使用 GNN 的改进模型 GCN 进行结构信息嵌入。嵌入模块利用的信息主要有两种，即结构信息和属性信息。交互模块的作用主要是将两个不同的知识图谱映射到同一向量空间，使得向量的计算在同一空间。目前联系两个知识图谱的桥梁主要是预对齐的实体对，通过预对齐的实体对在不同向量空间的转换和校准，统一两个知识图谱。对齐模块的作用主要是根据已经嵌入的实体向量来计算距离，此外，还能通过一些推理策略选择待对齐的实体。

值得注意的是，虽然基于知识表示学习的实体对齐方法取得了较为不错的效果，但是这并不意味着传统的实体对齐方法不具有研究价值。正如文献[5]也指出这两类方法是相辅相成的，结合起来考虑会有可能取得更好的效果。

随着知识图谱的不断完善，许多知识图谱都变得越来越复杂，规模也越来越大，原有的实体对齐算法需要进一步考虑执行效率和准确率。为了解决这个问题，并行处理技术受到了越来越多地关注。目前研究工作将并行处理技术应用到实体对齐任务中的是极少数[7]，有关大规模知识图谱的实体对齐问题仍然需要进行深入的研究和探索。

通过 4.3 节的对比分析，可以看到针对知识图谱结构信息的利用还有待于继续研究探索，无论是邻接实体还是实体间的关系，均对知识图谱的更准确表示起着至关重要的作用。使用神经网络嵌入知识图谱的结构信息时，如何缓解错误信息的传播至关重要。目前普遍使用高速门机制，使得错误传播的问题得到了一定程度的缓解，但是对于单跳和多跳实体的计算和信息传播仍需继续研究。

此外，在知识图谱结构信息嵌入表示方面，大多数实体对齐模型是以实体为中心，多方面信息辅助嵌入，在以后的研究中可以提高关系信息的占比，甚至可以以实体之间关系为中心研究嵌入表示，进而更深入地挖掘知识图谱的结构信息。除了结构信息，加入原知识图谱中的实体描述信息使得实体对齐效果显著提高，如 BERT-INT，甚至可以忽略结构信息。但是在真实大型知识图谱中，很多实体缺乏具体准确的描述信息，所以对结构信息以及其他未挖掘的信息有待于进一步深入研究。再者，在实体对齐任务中，大多数模型方法在通用数据集 DBP15K 数据集上获得了不错的效果。然而在实际大型真实知识图谱的表现一般，因此如何进一步提出不同种类的数据集也成为实体对齐领域的重要研究问题。

6 总结

本文给出了一个详细全面的实体对齐方法综述，详细深入地综述和比较了传统实体对齐方法和基于知识表示学习的实体对齐方法。首先对实体对齐的相关概念进行了介绍，之后对实体对齐数据集和评价指标进行了全面的概括。在此基础上，进一步详细深入地综述和比较了两大类实体对齐方法(即传统实体对齐方法和基于知识表示学习的实体对齐方法)。对每一类方法，进行了详细的划分和概括，并进行了对比分析。最后对实体对齐的未来研究方向进行了展望。目前，实体对齐任务越来越受到研究者们的关注，但是其中仍然存在着许多问题与不足，未来将会有更多实体对齐方法被提出，从而建立起多个知识图谱之间的链接，推动知识图谱领域的进一步发展。

在未来工作中，本文将进一步搭建实验基准模型，然后通过引入不同类别的信息（正如本文 4.2 节指出的结构信息、属性信息、实体描述信息等），通过实验数据进一步深入对比分析不同类别信息在实体对齐任务上的作用和实用场景。同时进一步进行基于注意力机制的实体对齐方法的实验对比分析。

参考文献

[1] Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. ACM Computing Surveys (CSUR), 2021, 54(4): 1-37.

[2] Zhao Xiaojuan, Jia Yan, Li Aiping, et al. A review of multi-source knowledge fusion technology . Journal of Yunnan University (NATURAL SCIENCE EDITION), 2020,42(03):459-473 (in Chinese).
(赵晓娟, 贾焰, 李爱平, et al. 多源知识融合技术研究综述. 云南大学学报(自然科学版), 2020,42(03):459-473)

[3] Wang Xuepeng, Liu Kang, he Shizhu, Liu Shulin, Zhang Yuanzhe, Zhao Jun. entity alignment algorithm for multi-source knowledge base based on network semantic tags .Chinese Journal of Computers,

2017,40 (3): 701-711(in Chinese).
(王雪鹏,刘康,何世柱,刘树林,张元哲,赵军.基于网络语义标签的多源知识图谱实体对齐算法.计算机学报,2017,40(3):701-711)

[4] Su Jialin, Wang yuanzhuo, Jin Xiaolong, Li manling, Cheng Xueqi. Entity alignment of knowledge map integrating semantic and structural information . Journal of Shanxi University (NATURAL SCIENCE EDITION), 2019,42 (01): 23-30(in Chinese).
(苏佳林,王元卓,靳小龙,李曼玲,程学旗.融合语义和结构信息的知识图谱实体对齐.山西大学学报(自然科学版),2019,42(01):23-30)

[5] Sun Z, Zhang Q, Hu W, et al. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs//Proceedings of the VLDB Endowment (VLDB), Tokyo, Japan, 2020: 2326-2340.

[6] Berrendorf M, Wacker L, Faerman E. A Critical Assessment of State-of-the-Art in Entity Alignment//European Conference on Information Retrieval. Lucca, Italy:Springer,2021: 18-32.

[7] Zhuang Yan, Li Guoliang, Feng Jianhua. Overview of entity alignment technology in knowledge base . Computer research and development, 2016,53 (01): 165-192(in Chinese).
(庄严,李国良,冯建华.知识图谱实体对齐技术综述.计算机研究与发展,2016,53(01):165-192)

[8] Meng Pengbo. Overview of entity alignment based on graph neural network . Modern computer, 2020 (09): 37-40(in Chinese).
(孟鹏博.基于图神经网络的实体对齐研究综述.现代计算机,2020(09):37-40)

[9] Zhao X, Zeng W, Tang J, et al. An experimental study of state-of-the-art entity alignment approaches. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(8): 1-14.

[10] Euzenat J, Ferrara A, Hollink L, et al. Results of the Ontology Alignment Evaluation Initiative 200// Proceedings of the 4th International Workshop on Ontology Matching (OM) collocated with the 8th International Semantic Web Conference (ISWC). Chantilly, USA, 2009: 1-56.

[11] Cohen W W, Richman J. Learning to match and cluster large high-dimensional data sets for data integration//Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining (KDD). Edmonton, Canada, 2002: 475-480.

[12] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning//Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining (KDD). Edmonton ,Canada, 2002: 269-278.

[13] Jean-Mary Y R, Shironoshita E P, Kabuka M R. Ontology matching with semantic verification. Journal of Web Semantics, 2009, 7(3): 235-251.

[14] Arasu A, Götz M, Kaushik R. On active learning of record matching packages//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD). Indianapolis, USA, 2010: 783-794.

[15] Suchanek F M, Abiteboul S, Senellart P. Paris: Probabilistic alignment of relations, instances, and schema//Proceedings of the VLDB Endowment, (VLDB).Seattle, USA, 2011: 157-168.

[16] Lacoste-Julien S, Palla K, Davies A, et al. Sigma: Simple greedy matching for aligning large knowledge bases//Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD).Chicago, USA, 2013: 572-580.

[17] Song D, Luo Y, Heflin J. Linking heterogeneous data in the semantic web using scalable and domain-independent candidate selection. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 143-156.

[18] Sun Z, Hu W, Li C. Cross-lingual entity alignment via joint attribute-preserving embedding//International Semantic Web Conference (ISWC). Vienna, Austria:Springer, 2017: 628-644.

[19] Trisedya B D, Qi J, Zhang R. Entity alignment between knowledge graphs using attribute embeddings//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).Honolulu, USA,2019: 297-304.

[20] Cai P, Li W, Feng Y, et al. Learning knowledge representation across knowledge graphs//Proceedings of the Workshops at the Thirty-First AAAI Conference on Artificial Intelligence(AAAI).San Francisco, USA,2017:704-710.

[21] Guo L, Sun Z, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019:1-13.

[22] Zeng Weixin, Zhao Xiang, Tang Jiuyang, et al. Iterative entity alignment based on reordering. Computer research and development, 2020, 57 (7): 1460-1471 (in Chinese).
(曾维新，赵翔，唐九阳，等. 基于重排序的迭代式实体对齐. 计算机研究与发展, 2020, 57(7): 1460-1471)

[23] Zhang Q, Sun Z, Hu W, et al. Multi-view knowledge graph embedding for entity alignment//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI).Macao, China, 2019: 5429-5435.

[24] Guan S, Jin X, Wang Y, et al. Self-learning and embedding based entity alignment. Knowledge and Information Systems, 2019, 59(2): 361-386.

[25] Zhu Jizhao, Qiao Jianzhong, Lin Shukuan. Entity alignment algorithm for representing learning knowledge map. Journal of Northeastern University (NATURAL SCIENCE EDITION), 2018, 39 (11): 1535-1539 (in Chinese).
(朱继召，乔建忠，林树宽. 表示学习知识图谱的实体对齐算法. 东北大学学报(自然科学版), 2018, 39(11):1535-1539)

[26] Zhu Q, Wei H, Sisman B, et al. Collective multi-type entity alignment between knowledge graphs//Proceedings of The Web Conference (WWW).Taipei,China,2020: 2241-2252.

[27] Su Jialin, Wang yuanzhuo, Jin Xiaolong, et al. Entity alignment method based on adaptive attribute selection. Journal of Shandong

University (Engineering Edition), 2020, 50 (1): 14-20 (in Chinese).
(苏佳林, 王元卓, 靳小龙, 等. 自适应属性选择的实体对齐方法. 山东大学学报(工学版), 2020, 50(1): 14-20)

[28] Kang Shize, Ji Lixin, Liu Shuxin, et al. A cross language entity alignment model based on entity description and knowledge vector similarity. Acta electronica Sinica, 2019, 47 (9): 1841-1847 (in Chinese).
(康世泽, 吉立新, 刘树新, 等. 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. 电子学报, 2019, 47(9): 1841-1847)

[29] Zhu H, Xie R, Liu Z, et al. Iterative Entity Alignment via Joint Knowledge Embeddings//Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI).Melbourne, Australia, 2017: 4258-4264.

[30] Ye R, Li X, Fang Y, et al. A Vectorized Relational Graph Convolutional Network for Multi-Relational Network Alignment//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI).Macao, China, 2019: 4135-4141.

[31] Yan Z, Peng R, Wang Y, et al. CTEA: Context and Topic Enhanced Entity Alignment for knowledge graphs. Neurocomputing, 2020, 410: 419-431.

[32] Halpin H, Hayes P J, Mccusker J P, et al. When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data//International Semantic Web Conference (ISWC). Berlin, Germany, 2010: 1-5.

[33] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems Man & Cybernetics, 2002, 21(3):660-674.

[34] Wang Q, Garrity G M, Tiedje J M, et al. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied & Environmental Microbiology, 2007, 73(16):5261-5267.

[35] Chih-Chung, Chang, Chih-Jen, et al. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems & Technology, 2011, 2(3): 1-27.

[36] Lin D. An Information-Theoretic Definition of Similarity//Proceedings of the Fifteenth International Conference on Machine Learning (ICML).Madison, USA, 1998: 296-304.

[37] Niwattanakul S, Singthongchai J, Naenudorn E, et al. Using of Jaccard Coefficient for Keywords Similarity//Proceedings of the Iaeng International Conference on Internet Computing & Web Services.Hong Kong, China,2013: 1-5.

[38] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2017, 29(12): 2724-2743.

[39] Liu Zhiyuan, sun Maosong, Lin Yankai, et al. Research progress of knowledge representation learning. Computer research and development, 2016, 53 (2): 247-261(in Chinese).
(刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展. 计算机研究与发展, 2016, 53(2): 247-261)

[40] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of theAdvances In Neural Information Processing Systems.Lake Tahoe, USA, 2013: 3111-3119.

[41] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. Computer Science, 2013:1-12.

[42] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data//Proceedings of theAdvances In Neural Information Processing Systems. Lake Tahoe, USA,2013: 2787-2795.

[43] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Québec, Canada,2014: 1112-1119.

[44] Lin Y, Liu Z, Sun M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion//Proceedings of the Twenty-ninthAAAI Conference on Artificial Intelligence (AAAI).Austin, USA,2015: 2181-2187.

[45] Xiao H, Huang M, Hao Y, et al. TransA: An adaptive approach for knowledge graph embedding//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Austin, USA,2015:1-7.

[46] Lin Y, Liu Z, Luan H, et al. Modeling relation paths for representation learning of knowledge bases//Proceedings of theConference on Empirical Methods in Natural Language Processing (EMNLP).Lisbon, Portugal, 2015: 1-10.

[47] Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases//Proceedings of theInternational Conference on Learning Representations (ICLR). San Diego, USA, 2015: 1-12.

[48] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction//Proceedings of the International Conference on Machine Learning (ICML), New York, USA, 2016: 1-12.

[49] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Phoenix, USA,2016: 1955-1961.

[50] Shi B, Weninger T. ProjE: Embedding projection for knowledge graph completion//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). San Francisco, USA,2017:1236-1242.

[51] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA,2018: 1811-1818.

[52] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks//Proceedings of the European Semantic Web Conference. (ESWC). Heraclion, Greece:Springer2018: 593-607.

[53] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the5th International Conference on Learning Representations (ICLR). Toulon, France, 2017: 1-14.

[54] Nguyen T D, Nguyen D Q, Phung D. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, USA,2018: 327-333.

[55] Jiang X, Wang Q, Wang B. Adaptive convolution for multi-relational learning//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis,USA,2019: 978-987.

[56] Vu T, Nguyen T D, Nguyen D Q, et al. A capsule network-based embedding model for knowledge graph completion and search personalization//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA,2019: 2180-2189.

[57] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules//Proceedings of the Advances in neural information processing systems. Long Beach, USA, 2017: 3856-3866.

[58] Guo L, Sun Z, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs//Proceedings of the International Conference on Machine Learning. (PMLR).Long Beach, USA, 2019: 2505-2514.

[59] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[60] Chen M, Tian Y, Yang M, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment//Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI). Melbourne, Australia, 2017: 1-7.

[61] Sun Z, Hu W, Zhang Q, et al. Bootstrapping Entity Alignment with Knowledge Graph Embedding//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden, 2018, 18: 4396-4402.

[62] Shi X, Xiao Y. Modeling multi-mapping relations for precise cross-lingual entity alignment//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).Hong Kong,China, 2019: 813-822.

[63] Cao Y, Liu Z, Li C, et al. Multi-Channel Graph Neural Network for Entity Alignment//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).Florence, Italy, 2019: 1452-1461.

[64] Li C, Cao Y, Hou L, et al. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China,2019: 2723-2732.

[65] Wong C M, Chen Q, Wu S, et al. Global Structure and Local Semantics-Preserved Embeddings for Entity Alignment//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence(IJCAI).Yokohama，Japan, 2020: 3658-3664.

[66] Sun Z, Huang J, Hu W, et al. TransEdge: Translating relation-contextualized embeddings for knowledge graphs//Proceedings of the International Semantic Web Conference. (ISWC).Auckland, New Zealand:Springer, 2019: 612-629.

[67] Sun Z, Wang C, Hu W, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 222-229.

[68] Mao X, Wang W, Xu H, et al. MRAEA: An Efficient and Robust Entity Alignment Approach for Cross-lingual Knowledge Graph//Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM). Houston, USA, 2020: 420-428.

[69] Chen J, Li Z, Zhao P, et al. Learning Short-Term Differences and Long-Term Dependencies for Entity Alignment//International Semantic Web Conference. (ISWC).Cancún, Mexico:Springer, 2020: 92-109.

[70] Teong K S, Soon L K, Su T T. Schema-Agnostic Entity Matching using Pre-trained Language Models//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. (ACM)Seattle,USA,2020: 2241-2244.

[71] Wang Zhichun, Lv Qingsong, et al. Cross-lingual knowledge graph alignment via graph convolutional networks//Proceedings of the 2018 Conference on Empirical Methods in Natural Language (EMNLP). Belgium, Brussels,2018: 349-357.

[72] Yang K, Liu S, Zhao J, et al. COTSAE: CO-Training of Structure and Attribute Embeddings for Entity Alignment//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 3025-3032.

[73] Liu Z, Cao Y, Pan L, et al. Exploring and Evaluating Attributes, Values, and Structure for Entity Alignment//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).Punta Cana, Dominican Republic, 2020: 6355-6364.

[74] Chen B, Zhang J, Tang X, et al. JarKA: Modeling Attribute Interactions for Cross-lingual Knowledge Alignment//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Delhi, India:Springer, 2020: 845-856.

[75] Wu Y, Liu X, Feng Y, et al. Jointly Learning Entity and Relation Representations for Entity Alignment//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).Hong Kong, China, 2019: 240-249.

[76] Wu Y, Liu X, Feng Y, et al. Relation-aware entity alignment for heterogeneous knowledge graphs//Proceedings of the 28th International Joint Conference on Artificial Intelligence (AAAI).Honolulu, USA, 2019: 5278-5284.

[77] Fey M, Lenssen J E, Morris C, et al. Deep Graph Matching Consensus//Proceedings of the International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia, 2020: 1-23.

[78] Wu Y, Liu X, Feng Y, et al. Neighborhood Matching Network for Entity Alignment//Proceedings of the 58th annual meeting of the Association for Computational Linguistics. The Association for Computational Linguistics (ACL). Seattle, USA, 2020:1-11.

[79] Mao X, Wang W, Xu H, et al. Relational Reflection Entity Alignment//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Seattle, USA,2020: 1095-1104.

[80] Zhu Y, Liu H, Wu Z, et al. Relation-Aware Neighborhood Matching Model for Entity Alignment//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 1-8.

[81] Zeng W, Zhao X, Wang W, et al. Degree-aware alignment for entities in tail//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA,2020: 811-820.

[82] Zeng W, Zhao X, Tang J, et al. Collective Entity Alignment via Adaptive Features//Proceedings of the 36th International Conference on Data Engineering (ICDE). Dallas, USA, 2020: 1870-1873.

[83] Yang H W, Zou Y, Shi P, et al. Aligning Cross-Lingual Entities with Multi-Aspect Information//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China,2019: 4422-4432.

[84] Wang Z, Yang J, Ye X. Knowledge graph alignment with entity-pair embedding//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic,2020: 1672-1680.

[85] Tang X, Zhang J, Chen B, et al. BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. (IJCAI). Yokohama，Japan, 2020: 3174-3180.

[86] Lin X, Yang H, Wu J, et al. Guiding cross-lingual entity alignment via adversarial knowledge embedding//Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China,2019: 429-438.

[87] Creswell A, Bharath A A. Inverting the generator of a generative adversarial network. IEEE transactions on neural networks and learning systems, 2018, 30(7): 1967-1974.

[88] Pei S, Yu L, Zhang X. Improving cross-lingual entity alignment via optimal transport//Proceedings of the 28th International Joint Conference on Artificial Intelligence (AAAI). Honolulu, USA,2019: 3231-3237.

[89] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport//Proceedings of the Advances In Neural Information Processing Systems. Lake Tahoe, USA,2013: 2292-2300.

[90] Pei S, Yu L, Hoehndorf R, et al. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference//Proceedings of the World Wide Web Conference (WWW).San Francisco, USA, 2019: 3130-3136.

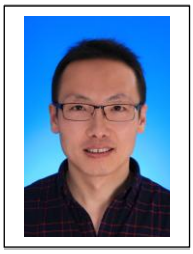
[91] Pei S, Yu L, Yu G, et al. REA: Robust Cross-lingual Entity Alignment Between Knowledge Graphs//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD).Macau, China, 2020: 2175-2184.

[92] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts//Proceedings of the 22nd International Conference on World Wide Web (WWW). Rio de Janeiro, Brazil, 2013: 1445-1456.

[93] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[94] Shi B, Weninger T. Open-world knowledge graph completion//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 1-8.

[95] Galárraga L, Razniewski S, Amarilli A, et al. Predicting completeness in knowledge bases//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM). Cambridge，UK,2017: 375-383.



ZhangFu, Ph.D., associate Professor.His current research interests include knowledge graphand Semantic Web.

YangLin-Yan,Ph.D.candidate.Her current research interest isentity alignment in knowledge graph.

LiJian-Wei,master student, his current research interest isentity alignment in knowledge graph.

Cheng Jing-Wei, Ph.D., lecturer.His current research interests include knowledge graphand Semantic Web.

Background

With the rapid development of knowledge graphs in recent years, a large number of knowledge graphs have emerged. However, there are serious heterogeneity and redundancy between knowledge graphs. Knowledge graph fusion aims to align and merge the heterogeneous and redundant information in the knowledge graph to form a global unified knowledge identification and association. Entity alignment is a key technology in the fusion process of knowledge graphs. The main purpose is to construct entity mappings between different knowledge graphs, which will refer to the same entities for matching.

In the early days, researchers used various characteristics of strings to perform entity alignment. In recent years, with the rapid development of knowledge representation learning technology, researchers have proposed many entity alignment methods based on knowledge representation learning, using deep learning to mine the structural information, attribute information, and description information of the knowledge graph. However, up to now, there is still a lack of comprehensive and in-depth methodological reviews on entity alignment technology.This article reviews and compares traditional entity alignment methods and entity alignment methods based on knowledge representation learning in detail. Aiming at traditional methods, the classification introduced entity alignment methods based on similarity calculation and relational reasoning, and in-depth study of the use of character features, attribute features, and relationship features in each type of method. At the same time, the advantages and disadvantages of different methods are deeply analyzed. Aiming at the entity alignment method based on knowledge representation learning, this article focuses on discussion, analysis and comparison. First of all, this paper abstracts this type of entity alignment method into a unified framework composed of three modules: embedded module, interaction module and alignment module, and gives a detailed overview of each method based on the three modules. Further, according to the different types of information used by the method, the existing methods are divided into eight types of methods based on structure information, attribute information, entity name information, entity description information, and comprehensive information, and each type of method is described in detail. Then, an in-depth comparative analysis of entity alignment methods based on knowledge representation learning is carried out. Finally, the main challenges of the entity alignment work are discussed, including the processing of sparse knowledge graphs, the lack of labeled data and noise issues, the efficiency of the method, etc., and the future of the work is prospected.