文章编号:1671-9352(2024)03-0071-10 DOI:10.6040/j.issn.1671-9352.1.2022.4484

基于图注意力神经网络的实体消歧方法

牛泽群[1],李晓戈[1,2,3*],强成宇[1],韩伟[1],姚怡[1],刘洋[3]
(1.西安邮电大学计算机学院,陕西 西安 710121;2.西安邮电大学陕西省网络数据分析与智能处理重点实验室,陕西 西安 710121;3.西安邮电大学西安市知识发现与应用工程技术中心,陕西 西安 710121)

摘要:针对链接对象为存在半结构化数据的知识库,提出了一种基于图注意力神经网络的短文本实体指称消歧方法。通过信息抽取与融入关键词,将含有半结构化数据的知识库构建为全局知识图谱;同时基于 Bert 预训练模型对短文本中的实体指称项进行嵌入融合;使用图注意力神经网络对全局知识图谱中候选实体节点进行加权聚合表征,并计算实体指称项与各候选实体之间的相似度得分,实现实体消歧。在 CCKS2019 数据集上的实验结果表明,基于图注意力神网络的实体消歧模型有效提高了实体消歧效果。

关键词:实体消歧;知识图谱;关键词提取;图注意力神经网络;自然语言处理
中图分类号:TP391 **文献标志码**:A
引用格式:牛泽群,李晓戈,强成宇,等. 基于图注意力神经网络的实体消歧方法[J]. 山东大学学报(理学版),2024,59(3):71-80,94.

Entity disambiguation method based on graph attention networks

NIU Zequn[1], LI Xiaoge[1,2,3*], QIANG Chengyu[1], HAN Wei[1], YAO Yi[1], LIU Yang[3]
(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China; 2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China; 3. Xi'an Knowledge Discovery and Application Engineering Technology Center, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China)

Abstract: We propose an entity disambiguation method based on graph attention networks for semi-structured knowledge base data. First, a global knowledge graph is constructed from the semi-structured knowledge base, and the entity reference items are embedded by Bert pre-trained model meanwhile. Next, graph attention networks which leverages masked self-attention layers is applyed on candidate entity nodes of global knowledge graph to fetch a vector of node level. Furtherly, we com pute similarity scores rank between the entity reference items and the candidate entity to complete the task of entity disambiguation. The experimental results on CCKS2019 dataset achieve state-of-the-art.

Key words: entity disambiguation; knowledge graph; keyword extraction; graph attention networks; natural language processing

0 引言

基于实体链接的实体消歧是解决短文本实体指称在知识库或知识图谱中存在同名实体一词多义或歧义的问题[1],例如在知识库中,"七里香"实体有着多种含义,如图 1 所示。目前,许多实体消歧方法是通过提取文本序列特征来实现实体消歧,但这类方法没有考虑到数据之间的相关性,而基于知识图谱的实体消歧方法可通过知识图谱的图形数据结构来表示实体之间的关系以及上下文特征,使数据之间的关联性得到扩充,

收稿日期:2022-09-29; **网络出版时间**:2023-12-05 14:52:02
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20231204.1019.002
基金项目:国家重点研发计划资助项目(2018YFB1402905);陕西省重点研发计划资助项目(2020GY-227);陕西省重点研发计划资助项目(2020ZDLGY09-05);陕西省技术创新引导专项基金(2022PT-49)
第一作者:牛泽群(1996—),男,硕士研究生,研究方向为自然语言处理. E-mail:1356903944@ qq.com
*通信作者:李晓戈(1962—),男,教授,博士,研究方向为自然语言处理. E-mail:lixg@ xupt.edu.cn

能更好地表示实体的特征属性[2-3]。例如图 2 所示,可根据知识图谱节点的上下文描述,将短文本中的“七里香”链接到知识图谱中的“七里香(周杰伦演唱的一首歌曲,由方文山作词,周杰伦谱曲,钟兴民编曲,收录在周杰伦 2004 年 8 月 3 日发行的同名专辑《七里香》中)”,从而消除知识库中其他义项所导致的歧义。

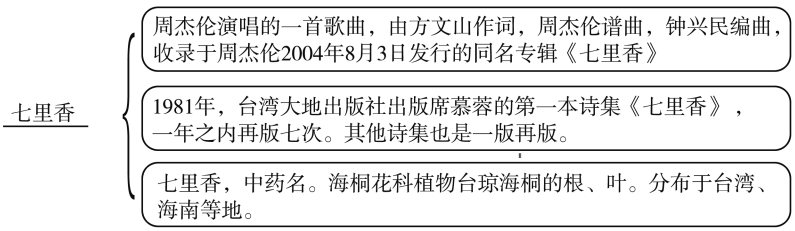

图 1 实体歧义性示例

Fig.1 Example of the entity ambiguity

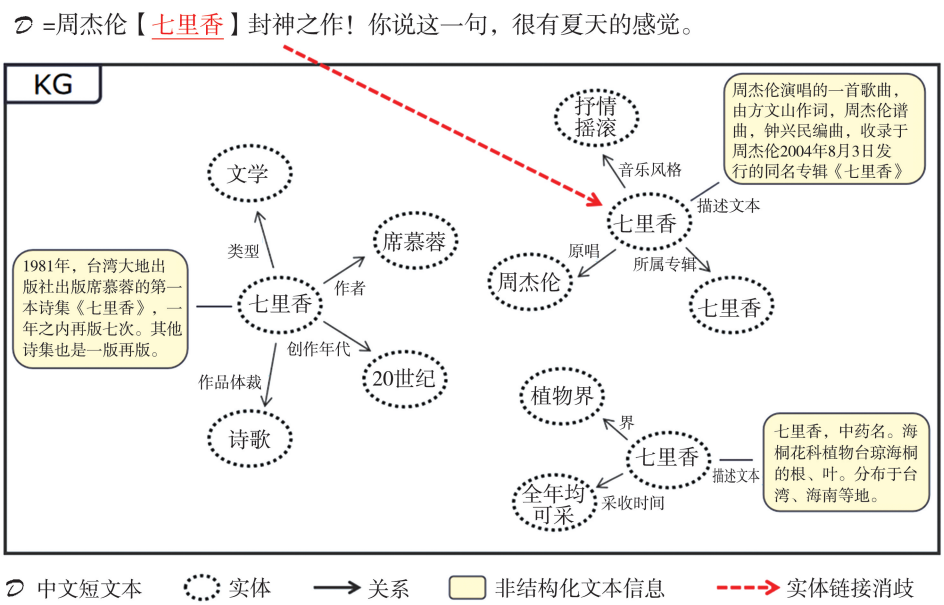

图 2 基于知识图谱实体链接消歧任务示例

Fig.2 Example of a disambiguation task linked based on knowledge graph entities

目前,多种知识库或知识图谱被应用于实体消歧,如 YAGO[4]、DBpedia[5]、Freebase[6]、百度百科、维基百科等,这些知识库包含了很多半结构化数据,其中同时包括了三元组与非结构化文本。通常研究者们在利用这些知识库数据进行实体消歧时,会将候选实体的所有描述信息进行拼接,提取该节点的上下文特征[7],但这种方法改变了原有的图形数据结构,忽略了半结构化数据的特殊性,没有考虑到不同邻居节点的知识描述对候选实体的权重影响,导致候选实体特征不能充分表示。

为了解决上述问题,本文针对利用含有半结构化数据的知识库作为链接对象的实体消歧任务,提出了一种基于全局知识图谱与图注意力神经网络的消歧模型,该模型改进了传统拼接处理半结构化数据的方式,通过信息抽取和融合关键词,将含有半结构化数据的知识库构建为全局知识图谱,并使用图注意力神经网络对候选实体节点进行信息表示,经实验证明,该消歧模型可以有效地提高消歧准确率。

本文的主要贡献:1)针对链接对象为存在半结构化数据的知识库,提出了一种基于图注意力神经网络的实体消歧模型,解决了实体消歧问题,有效地提高消歧准确率;2)通过将原知识库构建为全局知识图谱,建立了数据之间联系,减少了割裂子图的数量,提升了图的连通性与表示能力;3)通过与现有消歧模型进行实验对比,证明了引入 GAT 提取候选实体节点特征可有效提高消歧效果。

1 相关工作

实体消歧本质上是实体指称及其上下文与候选实体集合中实体的相似度排序问题,现有实体消歧方法

大体可划分为统计模型与多源知识方法、深度学习方法以及基于图的消歧方法[7]。

基于统计模型的实体消歧方法一般会统计实体相关的共现信息、主题词信息、关键词信息等特征作为排序的依据,实现实体消歧[8-11]。基于多源知识的实体消歧方法是利用特定的知识库提供丰富的背景信息来实现实体的消歧,通常情况下会采用融合多种来源的知识扩展候选实体集,避免采用单一知识库带来的局限性,从而提升实体消歧的准确性和全面性[12-14]。

基于深度学习的实体消歧方法无需手动构建特征,可通过神经网络自动对文本序列进行特征提取,为实体指称项、上下文信息、候选实体集合中的实体以及关联描述构建低维稠密空间下的语义表示,通过相似度排序选取最佳的目标实体。该方法表示了实体与实体之间的语义特征[7]。Francis-Landau 等[15]利用 CNN 来捕获实体指称上下文与目标实体的语义关系,从多个颗粒度的主题信息来衡量两者之间的语义相似性。Huang 等[16]利用深层神经网络与语义知识图所构成的新型深度语义相关性模型来衡量主题建模的语义相似性。

另外还有基于图的实体消歧方法,该方法首先会将文本中所有的实体构建为图,图中包含实体之间、候选实体之间、候选实体与实体指称项之间的关系,后续下游任务会在图的基础上提取特征,并根据相似度排序实现实体消歧。张涛等[17]基于维基百科的图结构,提出了新的语义关联度度量方法与学习排名框架,并利用随机游走衡量图中丰富的语义信息来处理实体排序任务。周金等[18]提出了一种基于图联合特征的实体消歧方法,联合主题、上下文、元数据等语义相似度,在经过扩充的图模型中利用随机游走和联合消歧实现消歧效果。

知识图谱是人工智能在知识表示和组织方面的最新技术,不仅为不同的实体之间建立了语义联系,还为信息检索、问答系统和推荐系统等工作提供了良好的数据支持。实体链接是找出一段文本中的实体指称,通过结合该实体的上下文与知识库或知识图谱中的语义信息,完成文本与知识图谱或知识库的的链接,该过程需要对每个文本中待链接的实体消除其所引用知识库或知识图谱中实体的歧义[19]。Mulang 等[20]利用维基百科进行实体消歧,通过将短文本、实体指称以及对应知识图谱中候选实体的多跳三元组进行拼接,使用 Transformer 进行特征提取,并将实体消歧问题转化为二分类问题。Cetoli 等[21]在维基百科知识图谱上获取大量(高达 1 500 个)的两跳三元组,并使用 RNN 对每个三元组拼接后的文本进行编码,将知识图谱作为神经网络消歧的额外信息来源。

对于知识图谱这种图形结构数据,需要利用图嵌入(graph embedding, GE)技术对其进行节点、节点属性以及关系的表征学习,最具有代表性的方法是基于深度学习的图嵌入方法[22]。Wang 等[23]提出了结构化深度网络嵌入模型,该模型具有多层非线性函数,可以捕获高度非线性的网络结构,并联合利用一阶和二阶相似度来获得最终的嵌入向量。Cao 等[24]提出了 DNGR 模型,通过结合随机漫游和深度自动编码器的方法捕获图结构信息,为每个顶点生成低维向量表示,避免了传统线性降维方法无法保持图非线性结构的问题。Hamilton 等[25]提出了基于图采样和聚合的图嵌入方法(GraphSAGE),该方法同时利用节点特征和结构信息得到 Graph Embedding 的映射,具有较强的扩展性。Kipf 等[26]提出了一种图卷积神经网络方法(GCN),该方法把频谱图卷积的定义进行简化,提高了计算效率。GCN 假设图中所有边的权重相同,这在现实应用场景中是不合理的,因此 Veličković等[27]提出了一种基于空间法的图注意力网络(GAT),将注意力机制引入图卷积中,其中注意力机制可在聚合邻居节点特征信息时确定节点的权重,同时图注意力网络使用了多头(multi-head)注意力机制,可学习不同子空间中的权重。本文采用图注意力神经网络模型在构建的知识图谱上学习图的连续性表示,提高实体链接消歧的准确率。

2　模型

将基于实体链接的实体消歧看为一个对短文本中实体指称项与知识图谱中多个候选实体的相似度排序问题,相似度得分最高的候选实体即为目标实体,通过五元组进行定义:

$$S=\{O,P,Q,T,R\},$$

式中:O 为实体指称项;P 是包含实体指称项的中文短文本,拥有实体指称项的上下文语义信息;Q 是知识图谱中的所有节点集合,包括候选实体与候选实体周围的邻居节点;T 是在知识图谱中与 O 同名的候选实体

集合;R 是候选实体的描述信息,本文指知识图谱中与候选实体节点相连的所有邻居节点集合;S 为实体消歧的定义。

本文提出的基于图注意力神经网络的实体消歧模型可以更好地适用于链接对象为含有半结构化数据的知识库,属于基于深度学习和图的混合消歧方法(见图 3)。其中文本特征表示是利用 BERT 预训练模型对短文本及文本中的实体指称项进行向量化表示,并进行特征融合;知识图谱构建与嵌入是将原知识库利用信息抽取技术构建为知识图谱并将每条数据的描述文本拼接利用 TF-IDF 抽取关键词,将抽取的关键词与原知识图谱融合,并采用图注意力网络对图谱节点进行特征表示;全连接层是将文本中特征融合后的实体指称项与经图注意力网络表征后的候选实体节点向量进行拼接,通过 Sigmoid激活函数得到候选实体的相似度得分,排序后取其最高分实体作为结果。

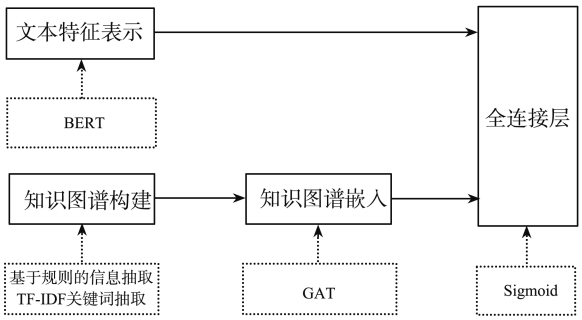

图 3 基于图注意力神经网络的实体消歧模型示意图

Fig.3 Schematic representation of the entity

2.1 文本特征表示

本文采用基于 BERT 预训练模型[28]的句子表示方法,对实体指称项以及包含实体指称项的中文短文本进行向量化表示,如图 4 所示。

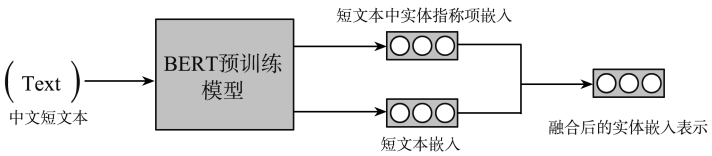

图 4 文本特征表示示意图

Fig.4 Text features represent a schematic diagram

BERT 采用 Transformer[29]作为主体框架并进行编码,预测时通过 Self-attention[30]双向综合考虑了上下文特征,并取高层的隐向量得到对应的句向量,得到的单条预测结果为 $1\times n$ 维矩阵,n 为特征维数,并将预测出的短文本向量与实体向量通过矩阵相加的方式进行特征融合,使其不仅包含了实体的语义信息,还结合了文本的上下文语义特征,将融合后的结果作为该实体指称项的特征表示。

2.2 知识图谱构建与嵌入

将知识库数据构建为知识图谱,并进行图嵌入后得到候选实体的向量表示。原数据中提供的是半结构化大型文本知识库,从中很难提取上下文特征。为了解决这个问题,本文将知识库数据构建为知识图谱,利用图表示实体之间的关系,以便获取候选实体的上下文特征。考虑到如果直接将知识库数据构建为知识图谱,各个候选实体之间会缺乏联系,关联性较少,图谱中几乎都是割裂的子图,没有充分利用到知识图谱的优点,因此本文采用融入关键词的方法,使各个候选实体子图之间产生联系,形成全连通图,构建全局知识图谱以提高下游图嵌入效果,见图 5。知识图谱构建与嵌入流程具体分为知识图谱构建、关键词补充、图嵌入表征这 3 步,见图 6。

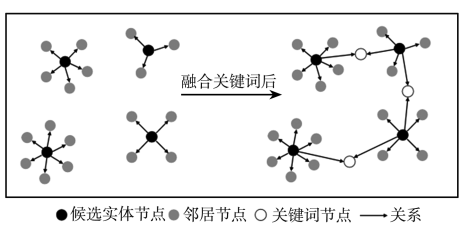

图 5 关键词融合示意图

Fig.5 Schematic diagram of the keyword fusion

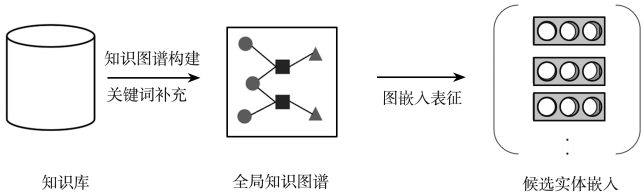

图 6 知识图谱构建与嵌入流程图

Fig.6 Knowledge graph construction and embedding flow chart

2.2.1　知识图谱构建

文本采用的知识库数据为 CCKS2019,其中某些节点属于半结构化文本数据,每条数据的类型主要包括候选实体、候选实体 ID、候选实体类型、别称、描述属性和属性值。根据上述内容,本文利用基于规则的信息抽取方法自上向下构建一个特殊的知识图谱模型,构建时采用预定义的方法预定义实体和关系,如表 1、2 所示。

表 1　实体预定义

Table 1　Entity predefined

实体类型	实体举例
候选实体(subject)	想你的夜
别称	Miss You Nights
属性值	女人如歌第四期
摘要文本	《想你的夜》是史丹丹的音乐作品,收录在《女人如歌第四期》专辑中。

表 2　关系预定义

Table 2　Relationships predefined

关系类型	三元组	三元组举例
别名	〈候选实体,别名,别称〉	〈想你的夜,别名,Miss You Nights〉
属性	〈候选实体,属性,属性值〉	〈想你的夜,所属专辑,女人如歌第四期〉
摘要	〈候选实体,摘要,摘要文本〉	〈想你的夜,摘要,摘要文本〉

因为属性是“摘要”的属性值,为长文本信息,所以将其定义为文本信息节点,其具体内容作为该节点属性。将候选实体节点作为中心节点;将别称、属性值和摘要文本节点作为邻居节点,用于中心节点的知识描述;将候选实体 ID、候选实体类型作为候选实体的节点属性。关键词融合如图 7 所示。

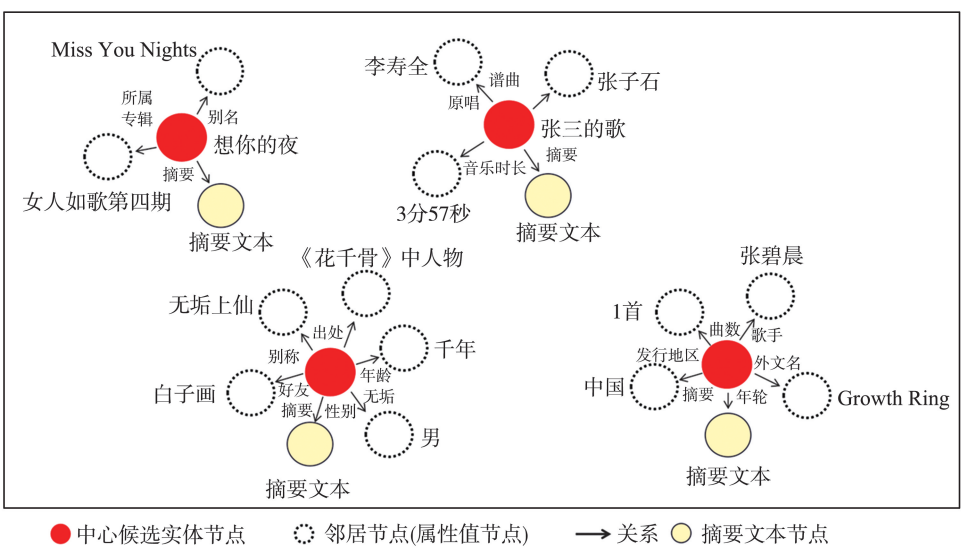

图 7　关键词融合示意图

Fig.7　Schematic representation of the knowledge graph model

2.2.2　关键词补充

为了减少图谱中割裂子图的存在,本文将原知识库中每个待消歧实体的所有属性、属性值和摘要文本进行拼接,构成该实体的知识描述长文本,并对所有实体对应的知识描述文本集合,利用 TF-IDF[31] 为每个待消歧实体抽取多个关键词,将提取的关键词属性作为该待消歧实体对应的特征节点,公式如下:

$$\mathrm{IDF}(x)=\log\frac{N+1}{N(x)+1}+1, \tag{1}$$

$$\text{TF-IDF}(x)=\mathrm{TF}(x)\cdot\mathrm{IDF}(x), \tag{2}$$

式中:N 表示待消歧实体的属性数量,x 表示待消歧节点的特征节点表示。在原本构建的图谱基础上,将抽取的所有关键词去重,并融入知识图谱,如果该实体属性值节点与关键词节点重复,则保留其一。新增预定义实体和关系如表 3、4 所示。

表 3 新增实体预定义
Table 3 New entity predefined

实体类型	实体举例
关键词	情歌

表 4 新增关系预定义
Table 4 New relation predefined

关系类型	三元组	三元组举例
Key word	〈候选实体,Key word,关键词〉	〈想你的夜,Key word,情歌〉

融合后的全局知识图谱在保留原图结构的基础上,通过引入关键词节点,使各条数据之间不再割裂,丰富了图谱中的关系,解决了割裂子图数据稀疏问题,并在后续知识图谱嵌入部分,通过多层图注意力神经网络的学习,中心实体节点可聚合多跳邻居节点的特征信息,增强节点表征能力,如图 8 所示。

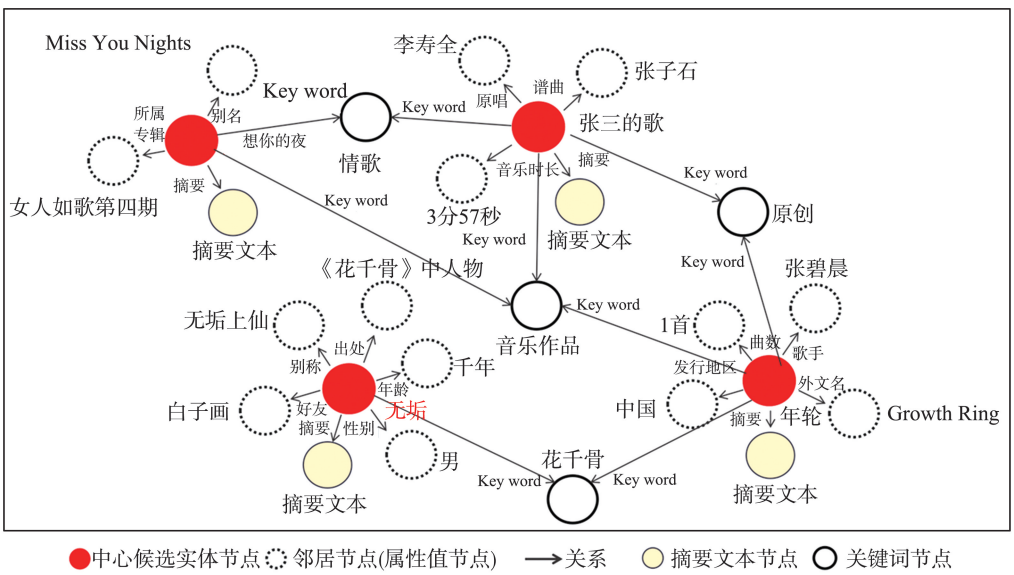

图 8 经关键词融合后全局知识图谱模型示意图
Fig.8 Schematic diagram of the global knowledge graph model after keyword fusion

2.2.3 图嵌入表征

Veličković等[27]提出的图注意力神经网络是利用注意力机制对邻居节点向量加权求和,根据不同邻节点的特征,为其分配不同的权值,并且邻近节点特征的权重完全取决于节点特征,独立于图结构,并在一系列基准数据集上的下游任务实验中证实图注意力神经网络可以提供很好的嵌入效果。据此,本文采用图注意力神经网络对知识图谱候选实体节点进行图嵌入节点表征。图嵌入表征结构如图 9 所示,图嵌入部分由输入层、特征提取层及图注意力神经网络层组成。

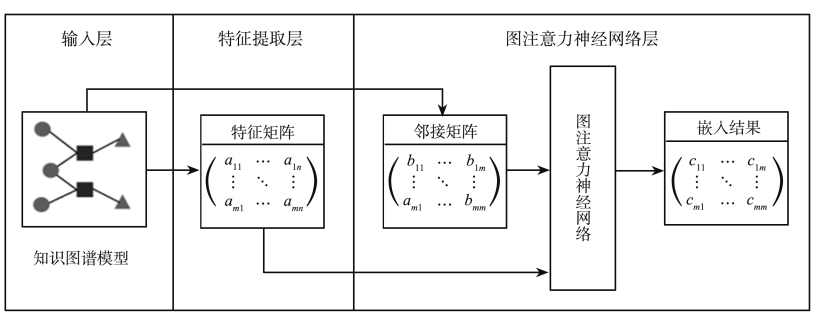

图 9 图嵌入表征结构
Fig.9 Graph embedding representation structure

2.2.3.1 输入层

图谱由 X 个节点 $V=\{v_1,v_2,\cdots,v_x\}$ 和 Y 条边 $E=\{e_1,e_2,\cdots,e_y\}$ 的集合组成。输入层一方面将上一部分构建的知识图谱模型输入到图注意力神经网络层;另一方面,直接从知识图谱中获取所有节点的文本信息,

输入到特征提取层,用于提取图谱节点的语义信息。

2.2.3.2　特征提取层

为了保持下游概率得分排序的可靠性与文本上下文语义的一致性,本文采取与上游文本特征表示部分相同的嵌入学习方式,采用同参数的 BERT 预训练模型对图谱中节点文本进行特征提取,得到 $m\times n$ 的特征矩阵,m 为图节点个数,n 为特征维数。

2.2.3.3　图注意力神经网络层

首先从输入层的知识图谱模型与特征提取层获得邻接矩阵和特征矩阵,然后将其输入到 2 层图注意神经网络进行图嵌入表示。该层输入是知识图谱和特征矩阵,输出是图注意力神经网络特征提取后的节点特征表示。

图注意力神经网络[27]通过引入注意力权重矩阵来学习图中任意相邻实体 x_i 和 x_j 的重要性,根据邻接矩阵判断节点间是否存在关系,邻接矩阵公式如下:

$$A=(a_{ij})+\boldsymbol{I}_N。\tag{3}$$

图注意力神经网络的更新机制为

$$\boldsymbol{h}_i^{l+1}=\sigma\Big(\sum_{j\in N_i}a_{ij}\boldsymbol{h}_i^l W^l\Big),\tag{4}$$

式中:$\boldsymbol{h}_i^{l+1}$ 和 $\boldsymbol{h}_i^l$ 分别为第 l+1 与 l 层 i 节点的向量表示,N_i 为 i 节点的邻居节点集合,a_{ij} 表示 i 和 j 节点之间的注意力相关系数矩阵,W^l 为第 l 层的参数矩阵,σ 为非线性激活函数。a_{ij} 的计算过程为

$$a_{ij}=\frac{\exp\Big(\text{LeakyReLU}\big(\partial\big[W\boldsymbol{h}_i\parallel W\boldsymbol{h}_j\big]\big)\Big)}{\sum_{k\in N_i}\exp\big(\text{LeakyReLU}\big(\partial\big[W\boldsymbol{h}_i\parallel W\boldsymbol{h}_k\big]\big)\big)},\tag{5}$$

式中:LeakyRelu 是激活函数;$\parallel$ 是将实体 x_i 和 x_j 的隐藏层向量进行拼接,并归一化得到注意力非对称权重矩阵,这种非对称性将图中节点的重要程度进行了区分。

注意力权重 a_{ij} 对邻居节点的隐藏层向量 $\boldsymbol{h}_i^l$ 加权平均,作为该节点的隐藏层表示 $\boldsymbol{h}_i^{l+1}$,如图 10 所示[27]。

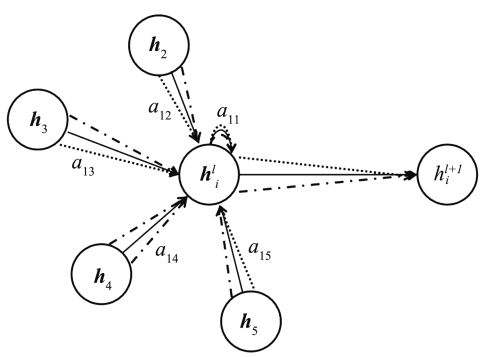

图 10　图注意力神经网络模型

Fig.10　GAT model

在完成图注意力神经网络层对图谱节点的特征更新后,输出表示同样为一个 $m\times n$ 的特征矩阵,m 为图节点个数,n 为特征维数。

2.3　全连接层

从嵌入后的节点特征矩阵中抽取知识图谱中所有候选实体的向量,记为 $\boldsymbol{A}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_p\}$,$a\in[1,p]$,单个候选实体向量记为 $\boldsymbol{T}_a$,同时将所有短文本特征融合后的实体指称项记为 $B=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_q\}$,$b\in[1,q]$,单个实体指称项记为 $\boldsymbol{T}_b$,将 $\boldsymbol{T}_a$ 与 $\boldsymbol{T}_b$ 拼接得到 T。

$$T=\text{concat}(\boldsymbol{T}_a,\boldsymbol{T}_b),\tag{6}$$

$$\text{Score}=\text{Sigmoid}(T)。\tag{7}$$

最后,通过全连接层使用 Sigmoid 激活函数为所有候选实体进行相似度打分,分数最高的候选实体即为目标实体。

3　实验

3.1　数据集与数据预处理

本文采用 CCKS2019 面向中文短文本实体链接任务所提供的数据集对模型进行验证,数据集中的内容主要来源于网络上的微博、词条、百科、新闻视频和文章的标题以及用户对话内容等,包括 90 000 条标注数据、39 925 条知识库待消歧候选实体知识信息,知识库属于半结构化数据,每条数据同时拥有三元组和长文本。

本文的任务是短文本中实体指称项与知识库中候选实体的消歧,需要对数据集进行分析和预处理。为了防止错误字符对下游消歧结果产生影响,本文对数据集中文本与知识库中相同字段的实体进行统一,防止

因存在标点符号、字母大小写、错别字等而无法匹配的问题。

3.2 评测标准

实验采用准确率(accuracy, A)和 F_1 值来评估消歧结果。

$$A=\frac{\text{num}_{\text{cor}}}{\text{num}_{\text{all}}} \tag{8}$$

式中:num_{cor}表示消歧结果中正确链接消歧的总数;num_{all}表示数据集总共需要消歧的样本总数;A 表示模型在数据集中正确链接消歧样本的百分比,A 值越高的模型性能越好。

$$P=\frac{\text{num1}}{\text{num2}}, \tag{9}$$

式中:P 表示精确度,num1 表示模型正确消歧的个数,num2 表示需要消歧的实体指称项总数。

$$R=\frac{\text{num1}}{\text{num3}}, \tag{10}$$

式中:R 表示召回率,num1 表示模型正确消歧的个数,num3 表示测试集中需要消歧的实体指称项总数。

$$F_1=\frac{2\times P\times R}{P+R}, \tag{11}$$

最终通过计算 F_1 分数判断模型的性能,F_1 值越高,模型性能越好。

3.3 对比实验设置

为了验证本文提出模型的效果,以及不同数据格式、注意力机制、图嵌入方式和图的连通性是否会对消歧结果产生影响,本文设置了基线模型对比实验和消融实验。

本文将 CCKS2019 中文短文本实体链接评测第一名的模型、张晟旗等[32]提出模型的实体消歧部分、BERT+GCN 模型以及 BERT+TF-IDF+GCN 模型作为本文的基线模型对比实验,将中文短文本实体链接评测第一名的模型与 BERT+GAT 模型作为消融实验。

评测第一名的模型是基于 BERT 实现实体消歧,将短文本以及待消歧实体的描述文本拼接,输入到 BERT 模型。取 CLS 位置向量输出,以及候选实体对应开始和结束位置的特征向量,3 个向量拼接,经过 sigmoid 激活函数得到候选实体的概率得分,排序选择得分最高的作为正确实体。张晟旗等[32]也是基于 BERT 模型,将待消歧实体与实体描述文本进行拼接,转换为长文本作为 BERT 输入,同时引入局部注意力解决长距离依赖问题,并强化局部的上下文信息。上述 2 种方法都是将半结构化数据拼接为非结构化文本作为数据载体,本文改进这 2 个实验模型输入的数据格式,将原本半结构化数据转化为结构化全局知识图谱,并在此基础上验证不同图神经网络和图连通性对消歧结果产生影响。基于知识图谱的对比模型如下。

① BERT+GCN:首先将短文本、实体指称项和知识图谱中的节点利用 BERT 进行编码,并融合短文本及实体指称项的特征作为实体指称项的表示,随后采用图卷积神经网络对知识图谱中待消歧实体节点进行嵌入表征,最后对实体指称项与待消歧实体集合进行相似度排序。

② BERT+GAT:在 BERT+GCN 的基础上,将图嵌入表示方式由图卷积神经网络换成图注意力神经网络,验证图注意力机制是否会提升实体消歧效果。

③ BERT+TF-IDF+GCN:在 BERT+GCN 的基础上,将提取的关键词融入知识图谱,验证当知识图谱连通性增强时能否提高消歧准确率。

3.4 实验结果与分析

实验将 90 000 条标注数据划分为 70 000 条训练集、10 000 条验证集以及 10 000 条测试集。利用图注意力神经网络进行图嵌入节点表征时,学习率设为 0.001,权重衰减设为 5×10^{-4},迭代次数设置为 100 次,dropout 设 0.3。消融实验结果如表 5 所示,基线模型对比实验结果如表 6 所示。

表 5 消融实验

Table 5 Ablation experiments

模型	A/%	P	R	F_1
BERT	77.7	0.76	0.78	0.77
BERT+GAT	72.2	0.71	0.72	0.71
BERT+Tf-Idf+GAT	81.3	0.83	0.82	0.82

表 6　基线模型对比实验结果
Table 6　Baseline model comparison experimental results

模型	A/%	P	R	F_1
BERT	77.7	0.76	0.78	0.77
BERT+Local Attention	78.5	0.77	0.78	0.77
BERT+GCN	69.8	0.71	0.69	0.70
BERT+Tf-Idf+GCN	76.0	0.77	0.78	0.77
BERT+Tf-Idf+GAT	81.3	0.83	0.82	0.82

表 5、6 的结果表明,本文提出的模型在实体消歧准确率上高于其他模型。BERT 模型与 BERT+Local Attention模型采用的是非结构化文本作为数据载体,这会导致知识库中各候选实体及描述文本数据相互独立,使候选实体上下文提取特征不足,从而影响下游消歧效果。BERT+GCN 与 BERT+GAT 模型虽然利用了知识图谱作为数据载体,但图谱中各个候选实体之间缺乏联系,几乎都为割裂的子图,没有有效地利用到知识图谱的优点。BERT+Tf-Idf+GCN 模型通过增加关键词补充了知识图谱实体间的关系,但采用图卷积神经网络进行图嵌入表示,假设节点之间的连接关系重要度一样,是对中心节点的平均聚合表征,没有引入注意力机制。本文将存在半结构化数据的知识库转化为全局知识图谱,在各候选实体数据之间建立了联系,表示中心实体时,可融合多跳周围邻居节点信息,在特征提取效果上优于传统文本表示方式,从而有效提高了实体消歧义准确率。

4　结束语

针对存在半结构化数据的知识库,本文提出的基于图注意力神经网络的实体消歧方法,通过将半结构化知识库构建为知识图谱,使实体与描述文本之间产生关系;利用抽取的关键词融合进图,形成全局知识图谱,使知识图谱中各个候选实体之间不再割裂,丰富了图谱中的关系;使用多层图注意力神经网络加权聚合邻居节点信息到中心实体的特性,强化了候选实体节点的上下文信息,增强了知识图谱的连续性表示,能更好地进行学习表征。实验结果表明,本文的模型有效地提高了实体消歧准确率。在未来的工作中,将进一步研究图神经网络,并引入关系的注意力权重来提高模型的准确率。

参考文献:

[1] 李天然,刘明童,张玉洁,等.基于深度学习的实体链接研究综述[J]. 北京大学学报(自然科学版), 2021, 57(1):91-98.
LI Ziran, LIU Mingtong, ZHANG Yujie, et al. Review of entity linking research based on deep learning[J]. Journal of Peking University(Natural Science Edition), 2021, 57(1):91-98.

[2] 刘峤,李杨,段宏,等.知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3):582-600.
LIU Qiao, LI Yang, DUAN Hong, et al. Overview of knowledge graph construction technology[J]. Computer Research and Development, 2016, 53(3):582-600.

[3] 段宗涛,李菲,陈柘.实体消歧综述[J]. 控制与决策, 2021, 36(5):1025-1039.
DUAN Zongtao, LI Fei, CHEN Zhe. Overview of entity disambiguation[J]. Control and Decision, 2021, 36(5):1025-1039.

[4] BOIŃSKI T, SZYMAŃSKI J, DUDEK B, et al. NLP questions answering using DBpedia and YAGO[J]. Vietnam Journal of Computer Science, 2020(3):1-16.

[5] SINGH K, LYTRA I, RADHAKRISHNA S A, et al. No one is perfect:analysing the performance of question answering components over the DBpedia knowledge graph[J]. Journal of Web Semantics, 2020, 65(1):100594.

[6] HUANG Zhipeng, BOGDAN C, REYNOLD C, et al. Entity-based query recommendation for long-tail queries[J]. ACM Transactions on Knowledge Discovery from Data, 2018, 12(6):1-24.

[7] 张丹阳,李楠,陈翀.实体链接技术研究述评[J]. 情报工程, 2020, 6(6):45-55.
ZHANG Danyang, LI Nan, CHEN Chong. Review of research on entity link technology[J]. Information Engineering, 2020, 6(6):45-55.

[8] HUANG D, WANG J. An approach on Chinese microblog entity linking combining baidu encyclopaedia and word2vec[J]. Procedia Computer Science, 2017, 111:37-45.

[9] GUO Zhaochen, BARBOSA Denilson. Robust named entity disambiguation with random walks[J]. Semantic Web, 2017, 9(11):1-21.

[10] 武川,陆伟.基于上下文特征的短文本实体链接研究[J]. 情报科学, 2016(2):144-147.
WU Chuan, LU Wei. Research on short text entity linking based on context features[J]. Intelligence Science, 2016(2):144-147.

[11] 谭咏梅,王睿,李茂林.基于上下文信息和排序学习的实体链接方法[J]. 北京邮电大学学报, 2015(5):33-36.
TAN Yongmei, WANG Rui, LI Maolin. Entity linking method based on context information and ranking learning[J]. Journal of Beijing University of Posts and Telecommunications, 2015(5):33-36.

[12] 昝红英,吴泳钢,贾玉祥,等.基于多源知识的中文微博命名实体链接[J]. 山东大学学报(理学版), 2015(7):9-16.
ZAN Hongying, WU Yonggang, JIA Yuxiang, et al. Chinese Weibo named entity link based on multi-source knowledge[J]. Journal of Shandong University(Natural Science), 2015(7):9-16.

[13] 周鹏程,武川,陆伟.基于多知识库的短文本实体链接方法研究: 以 Wikipedia 和 Freebase 为例[J]. 现代图书情报技术, 2016(6):1-11.
ZHOU Pengcheng, WU Chuan, LU Wei. Research on short text entity linking method based on multiple knowledge bases: taking Wikipedia and Freebase as examples[J]. Modern Library and Information Technology, 2016(6):1-11.

[14] SUN Chenchen, SHEN Derong, KOU Yue, et al. Topological features based entity disambiguation[J]. Journal of Computer Science and Technology, 2016, 31(5):1053-1068.

[15] FRANCIS-LANDAU M, DURRETT G, KLEIN D. Capturing semantic similarity for entity linking with convolutional neural networks[J/OL]. arXiv, 2016. https://arxiv.org/pdf/1604.00734.pdf.

[16] HUANG H, HECK L, JI H. Leveraging deep neural networks and knowledge graphs for entity disambiguation[J]. CoRR, 2015, abs/1504.07678.

[17] 张涛,刘康,赵军.一种基于图模型的维基概念相似度计算方法及其在实体链接系统中的应用[J]. 中文信息学报, 2015, 29(2):58-67.
ZHANG Tao, LIU Kang, ZHAO Jun. A graph model based Wiki concept similarity calculation method and its application in entity link systems[J]. Chinese Journal of Information Technology, 2015, 29(2):58-67.

[18] 周金,朱永华,张铁男,等.基于图的联合特征实体链接方法[J]. 上海大学学报(自然科学版), 2020, 26(5):747-755.
ZHOU Jin, ZHU Yonghua, ZHANG Tienan, et al. Graph based joint feature entity linking method[J]. Journal of Shanghai University(Natural Science), 2020, 26(5):747-755.

[19] 郭宇航,秦兵,刘挺,等.实体链指技术研究进展[J]. 智能计算机与应用, 2014, 4(5):9-13.
GUO Yuhang, QIN Bing, LIU Ting, et al. Research progress in physical chain finger technology[J]. Intelligent Computers and Applications, 2014, 4(5):9-13.

[20] MULANG' I O, SINGH K, PRABHU C, et al. Evaluating the impact of knowledge graph context on entity disambiguation models[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020:2157-2160.

[21] CETOLI A, BRAGAGLIA S, O' HARNEY A D, et al. A neural approach to entity linking on wikidata[C]//European Conference on Information Retrieval. Berlin:Springer, 2019:78-86.

[22] 祁志卫,王笳辉,岳昆,等.图嵌入方法与应用:研究综述[J]. 电子学报, 2020, 48(4):808-818.
QI Zhiwei, WANG Jiahui, YUE Kun, et al. Graph embedding methods and applications:research review[J] Journal of Electronics, 2020, 48(4):808-818.

[23] WANG D, CUI P, ZHU W. Structural deep network embedding[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016:1225-1234.

[24] CAO S, LU W, XU Q. Deep neural networks for learning graph representations[C]//Proceedings of the AAAI Conference on Artificial Intelligence.[S.l.]: AAAI, 2016, 30(1).

[25] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[J/OL]. arXiv, 2017. https://arxiv.org/pdf/1706.02216.pdf.

[26] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J/OL]. arXiv, 2016. https://arxiv.org/abs/1609.02907.

(下转第 94 页)