基于关系指数和表示学习的领域集成实体链接

蒋胜臣 [1, 2] 王红斌 [1, 2] 余正涛 [1, 2] 线岩团 [1, 2] 王红涛 [1, 2]

摘 要 本文针对现有方法不能很好结合文本信息和知识库信息的问题, 提出一种基于关系指数和表示学习的领域集成实体链接方法.首先, 本文构建了特定领域知识库; 其次, 运用表示学习从文本信息中得到的向量表示计算实体指称项的上下文、主题关键词、扩展词三个特征的相似度; 然后, 利用知识库中的关系信息计算候选实体的关系指数; 最后, 将这三种相似度及关系指数相融合, 用于实体链接. 实验结果表明, 相较于现有方法, 本文方法能够有效地提高 F1 值, 并且该方法不需要标注语料, 更加简单高效, 适应于缺少标注语料的特定领域.

关键词 集成实体链接, 特定领域, 表示学习, 关系指数

引用格式 蒋胜臣, 王红斌, 余正涛, 线岩团, 王红涛. 基于关系指数和表示学习的领域集成实体链接. 自动化学报, 2021, **47**(10): 2376−2385

DOI 10.16383/j.aas.c180705

Domain Integrated-Entity Links Based on Relationship Indices and Representation Learning

JIANG Sheng-Chen[1, 2] WANG Hong-Bin[1, 2] YU Zheng-Tao[1, 2] XIAN Yan-Tuan[1, 2] WANG Hong-Tao[1, 2]

Abstract Aiming at the problem that the existing methods can′ t combine text information and knowledge base information well, this paper proposes a domain integrated-entity links based on relationship indices and representation learning. Firstly, this paper builds a domain-specific knowledge base; Secondly, using the vector representation of learning representation from the text information to calculate the similarity of the three features of the context, topic keywords and extension words of the entity referential item; Then using the relationship information in the knowledge base to calculate the relationship index of the candidate entities; Finally, these three similarities and relationship indices are combined for physical links; The experimental results show that compared with the existing methods, the proposed method can effectively improve the F1 value, and the method does not need to label the corpus, which is simpler and more efficient, and is suitable for the specific field lacking the labeled corpus.

Key words Integrated-entity link, specific field, representation learning, relationship index

Citation Jiang Sheng-Chen, Wang Hong-Bin, Yu Zheng-Tao, Xian Yan-Tuan, Wang Hong-Tao. Domain integrated-entity links based on relationship indices and representation learning. *Acta Automatica Sinica*, 2021, **47**(10): 2376−2385

实体链接是指将文本中存在歧义的实体正确链接到知识库中无歧义的候选实体的过程[1−2], 实体链接的相关研究有助于知识库的自动填充[3], 也有助于信息检索的研究[4], 同时实体链接与跨文本指代消解、词义消岐, 实体消岐等诸多自然语言研究领域有着紧密联系. 目前关于实体链接的研究方法, 主要思想是通过计算实体指称项与其候选实体的多种特征相似度, 选择知识库中无歧义实体进行链接. 早期研究以单实体为对象, Bunescu[5] 和 Ganea 等[6] 使用词袋模型计算指称项与候选实体的相似度, 选取相似度最高的候选实体作为目标实体; Cucerzan[7] 和 Nguyen 等[8] 通过维基百科页面锚文本、重定向页面等信息计算指称项与候选实体的相似度; Zeng[9] 利用第三方知识库对候选实体特征进行扩充使得实体链接准确率提高. 以单实体为对象的实体链接方法忽略了文本中共现实体间的语义关系, 并且计算效率不高.针对以上问题, 研究者们结合已有知识库中存在的信息, 提出以集成实体作为对象进行实体链接的集成实体链接方法. Han 等[10] 通过构建候选实体语义相关图进行集成实体链接; Liu 等[11] 提出基于图的集成实体链接方法, 以实体指称项和

收稿日期 2019-05-24 录用日期 2019-07-17
Manuscript received May 24, 2019; accepted July 17, 2019
国家自然科学基金 (61562052, 61462054)
Supported by National Natural Science Foundation of China (61562052, 61462054)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 昆明理工大学信息工程与自动化学院 昆明 650500 2. 昆明理工大学云南省人工智能重点实验室 昆明 650500
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500

候选实体作为顶点构建有向图, 通过计算出入度和语义相似度进行集成实体链接; Ferragina 等[12] 引入了概率化链接的思想, 提出了一个面向短文本的集成实体链接算法. 这些研究在一定程度弥补了单实体链接忽视共现实体间语义相关性的不足, 但是却在一定程度上忽略了指称项本身具有的文本特征, 对文本信息利用率不高.

近些年随着深度学习在自然语言中的应用, 利用表示学习计算语义相似度成为一种新的思路[13−14]. 随着 Bengio 等[15] 提出表示学习模型, 通过表示学习表征实体深层语义信息计算相似度成为实体链接任务的新趋势[16−17]. Mikolov 等[18] 和 Goldberg[19] 对向量空间中词表示的有效嵌入进行了评估; Kar 等[20] 将表示学习用于特定任务领域的实体消歧; Moreno 等[21] 等通过扩充锚文本对文本中的单词和知识库中的实体进行联合学习得到相应的向量表示形式, 从而进行实体链接.

以上研究都是在通用领域, 其有丰富的通用语料和消歧特征[22]; 而对于特定领域, 往往存在语料不足, 另外流行度等消歧特征不明显的问题, 针对这些问题, 本文提出了一种新的基于关系指数和表示学习的领域集成实体链接方法. 首先, 构建特定领域知识库, 以作为实体链接的基础; 其次, 通过 LDA 主题模型、word2vec 模型和 TransE 模型训练本文收集到的背景语料和特定领域知识库中的三元组, 得到蕴含知识和主题信息的实体指称项和候选实体的向量表示; 再利用得到的向量表示和 LDA 主题模型抽取实体指称项所在主题的领域关键词; 然后, 结合词扩展, 得到实体指称项的扩展词; 再利用得到的特征, 计算指称项与候选实体的上下文、领域关键字、扩展词三种特征相似度; 同时利用知识库中丰富的关系信息, 得到候选实体的关系指数; 最后, 将三种特征相似度和关系指数相融合, 得到最后的相似度. 本文的主要贡献主要有: 1) 利用表示学习, 同时将文本词向量表示和知识库的知识表示嵌入到同一个语义空间, 融合了文本信息和知识库信息; 2) 收集了语料, 获取了特定领域相关知识, 构建了特定领域知识库; 3) 将关系属性融入到实体链接中, 实现了实体的语义属性和关系属性的融合.

1　研究方法

本文提出的方法具体步骤是: 首先, 构建特定领域知识库, 以作为实体链接的基础; 其次, 通过 LDA 主题模型、word2vec 模型和 TransE 模型训练本文收集到的背景语料和特定领域知识库中的三元组, 得到蕴含知识信息和主题信息的实体指称项和候选实体的向量表示; 再利用得到的向量表示和 LDA 主题模型抽取实体指称项所在主题的领域关键词; 其次, 结合词扩展, 得到实体指称项的扩展词; 然后, 利用得到的特征, 计算指称项与候选实体的上下文、领域关键字、扩展词三种特征相似度; 同时利用知识库中丰富的关系信息, 得到候选实体的关系指数; 最后, 将三种特征相似度和关系指数相融合, 得到最终相似度. 将相似度最高的候选实体作为最终链接对象.

本文方法包括 5 部分: 特定领域知识库构建、融合知识和主题信息的词向量训练、候选实体的生成、多特征生成、实体链接. 如图 1 所示.

1.1　领域知识库构建

本文针对特定领域, 在分析领域属性的基础上, 通过人工定义知识体系, 从百度百科等网站上收集了相关语料, 包括旅游景点语料、野生菌语料、茶叶语料、中国少数民族语料, 小吃语料和药材语料, 交通方式和住宿信息语料共计 96 674 个词条, 构建了具有一定规模的特定领域知识库. 然后将识别好的领域实体和实体间关系采用批量导入的方式导入到图数据库 Neo4j 进行管理. 本文使用自构建的特定领域知识库作为实体链接任务的支撑, 并结合百度百科作为第三方知识库对自构建的特定领域知识库中的实体属性进行有效补充. 具体方法是针对知识库中的每个实体, 通过它在百度百科相应的概念页面, 抓取页面中 Infobox 的半结构化三元组数据. 然后利用 Neo4j 图数据库进行管理. 对本地知识库中同名实体采用加后缀标签的方式进行区分, 且后缀标签用小括号与实体隔离. 例如: 实体 “香格里拉”. 在本地知识库中有三个相应实体, 分别加上后缀标签 “地名”、“酒店”、“电影”, 并用小括号进行隔离. 如: 香格里拉 (酒店)、香格里拉 (地名)、香格里拉 (电影).

1.2　融合知识信息和主题信息的词向量模型训练

1.2.1　主题关键词特征提取

特定领域的实体链接可以利用领域特征进行实体链接[23−24], 领域关键词表征了领域的主要语义信息和领域特征, 但是基于领域关键词的相似度计算主要是从全局上下文信息出发, 并没有考虑到文本局部的上下文信息, 针对这个问题, 本文提出利用 LDA 主题模型对训练语料上下文进行主题分类, 通过在不同主题下对多义词与主题词结合进行语义扩充, 计算词与词之间的余弦相似度进行 K-Means 聚类, 选择离聚类中心最近的 m 个词作为主题关键词.

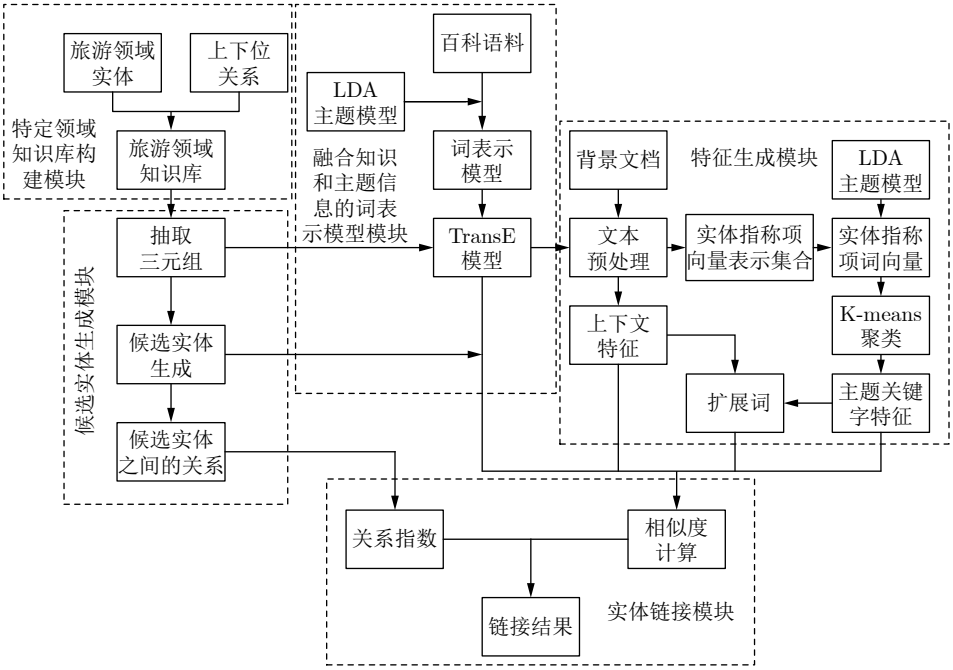


图 1 模型框架图

Fig. 1 Frame diagram of the model

1.2.2 融合主题信息的词向量模型训练

Mikolov 等[18] 提出 Word2vec, 通过神经网络将词表示在一个低维稠密的向量空间中, 利用距离和角度反映出词语之间的语义信息; 本文选择 Google 的开源工具包 word2vec, 采用 Skip-gram 模型作为词向量训练的基本模型, 其主要思想为根据中心词最大概率得到出其上下文:

$$L=\frac{1}{K}\sum_{k=1}^{K}\sum_{-N\le j\le N, j\neq 0}\ln p(w_{k+j}|w_k) \tag{1}$$

其中, w_k 是中心词, w_{k+j} 表示中心词的上下文, N 是训练时窗口的大小, 在本文中没有对窗口设置对比实验, 按照实验经验, 设窗口大小为 5. $p(w_{k+j}|w_k)$ 表示在中心词 w_k 的条件下, w_{k+j} 生成的概率, 利用 softmax 函数求得:

$$p(w_{k+j}|w_k)=\frac{\exp\left(\left(v^*_{w_{k+j}}\right)^{\mathrm{T}}v_{w_k}\right)}{\sum\limits_{w=1}^{n}\exp\left(\left(v^*_{w}\right)^{\mathrm{T}}v_{w_k}\right)} \tag{2}$$

其中, $v^*_{w_{k+j}}$ 和 v_{w_k} 是输入、输出的潜在变量, n 表示词汇表的大小. 主题信息是词语信息的重要组成部分, 对词语语义理解、文章类别分析都有重要意义. 为了将主题词信息融入词向量表示中, 本文利用第 2.2.1 节得到每个实体的主题关键词, 计算实体与其主题关键词的距离:

$$D_g=\frac{\sum\limits_{j=1}^{m}d(w_k,w_j)}{m} \tag{3}$$

其中, $d(w_k,w_j)$ 表示词 w_k 和 w_j 向量表示的欧几里得距离, m 表示词 w_k 的主题词个数. 将主题信息融入词向量表示中:

$$J_g=\alpha D_g-L \tag{4}$$

其中, α 为权重值, 我们的目标是最小化 J_g, 通过将主题关键词的距离融入词向量表达中, 使得同主题词之间的向量表示更接近. 对没有同主题关键词的词语, 直接按照 Skip-gram 模型训练出其向量表示形式. 通过对训练出的词向量与同主题词计算相似度并参考 Xu 等[25] 的实验参数, 设置 $\alpha=0.8; m=6$.

1.2.3 TransE 模型的联合学习

Bordes 等在 Mikolov 的 word2vec 词表示学习模型的基础上提出了 TransE 模型[26], 将知识库中的关系看作实体间的某种平移向量. 通过 TransE 模型对构建的特定领域知识库中的三元组进行训练, 得到知识库中实体和关系的向量表示. 针对现有的实体链接方法, 无法将知识库信息和文本信息更好的融合, 造成在实体链接中无法利用更多的文本信息和知识库信息, 在本文中, 为了将知识库信息与文本信息融合, 以达到更佳的实体链接效果, 我们将第 2.2.2 节中融合主题信息的词向量表示与知识表示模型 TransE 联合学习. 首先利用收集到

的三元组语料预训练 TransE 模型, 得到实体与关系的向量表示, 再将第 2.2.2 节得到的融合主题信息的词向量表示形式, 替换原有的实体向量表示, 计算两者的尾实体的距离:

$$D_z = \sum_{i=1}^{n} d(w_{k,r}, w^*_{k,r}) \tag{5}$$

其中, $\boldsymbol{w}_{k,r}$ 表示 TransE 模型得到的原实体 w_k 和关系 r 的向量之和, $\boldsymbol{w}^*_{k,r}$ 表示 w_k 在融合主题信息的词向量模型中的向量表示和关系 r 的向量之和, n 表示实体个数. 通过最小化 D_z, 使得词向量表示和知识表示相互约束训练模型, 最终得到融合结构知识的词向量表示.对于在自构建的本地知识库中没有实体相对应的词语, 将它们输入到训练好的模型中得到新的向量. 我们称之为融合伪知识的词向量表示, 这样做是将文本中的词与自构建本地领域知识库中实体向量表示嵌入到同一个语义空间中, 达到融合文本信息和知识库信息的目的, 也为后面的相似度计算提供方便. 本文没有对 TransE 模型的参数对实验结果的影响做特定实验, 向量维数设为 200, 边缘超参数设为 1, 学习速率设为 0.0001, 选用 $L2$ 作为距离计算公式. 在整个融合知识和主题信息的词向量表示过程中, 向量维度统一设为 200, 整体模型框架图如图 2 所示.

1.3　候选实体生成

1.3.1　候选实体的选取

对于候选实体的生成, 首先要识别出文本中所有的实体指称项, 将实体指称项组成集合 $M = \{m_1, m_2, \cdots, m_n\}$, 其中 n 表示文本中实体指称项的个数. 然后针对每个实体指称项 m_i, 在自构建的特定领域知识库中寻找与之同名实体 (不包括括号内的实体后缀标签) 并组合成集合, 作为它的候选实体集合 $N_i = \{n_{i1}, n_{i2}, \cdots\}$. 如果知识库中没有同名实体, 则把相应的实体指称项归为空实体; 当候选实体个数小于等于 4 时, 选取指称项所有的候选实体作为它最终的候选实体; 当候选实体个数大于 4 时, 计算指称项与候选实体的上下文相似度, 选取相似度最大的 4 个候选实体作为最终的候选实体. 上下文相似度计算公式为:

$$S_w = \frac{\sum_{i=1}^{d}\sum_{h=1}^{u} \cos(E(G_i), E(G^*_h))}{d \cdot u} \tag{6}$$

其中, $\boldsymbol{E}(G_i)$ 和 $\boldsymbol{E}(G^*_h)$ 分别表示实体指称项的上下文词和其候选实体直接三元组尾实体的向量表示; d 和 u 分别表示实体指称项的上下文词的个数和其候选实体直接三元组尾实体的个数.

1.3.2　候选实体关系属性的计算

针对集成实体链接, 关系属性是候选实体的重要属性之一, 基于实体指称项语义相近, 则它们在知识库中的无歧义实体也应该具有关系的思想. 例如: 实体指称项 “香格里拉”和 “丽江”, 它们语义相近, 则它们在知识库中的无歧义实体 “香格里拉 (旅游胜地)”和 “丽江 (旅游胜地)”也具有相应的关系. 本文将候选实体的关系属性分为直接关系属性和间接关系属性. 1) 直接关系属性计算自构建的特定领域知识库中含有丰富的关系属性, 根据第 2.3.1 生成文本中实体指称项的候选实体集合 $H = \{N_1, N_2, \cdots, N_n\}$, 其 N_i 表示第 i 个实体指称项的候选实体集合, n 为背景文档中实体指称项个数. 结合自构建的领域知识库, 得到候选实体的直接关系属性, 具体方法为: 对候选实体集合 N_i 中的每个元素分别与其他 $n-1$ 个候选实体集合中的每个元素进行关系查找, 如果两者之间存在直接三元组, 则两个元素之间的关系指数为 1, 不存在则关系指数为 0. 对于第 i 个实体指称项的第 j 个候选实体 n_{ij} 的直接关系指数, 计算公式为:

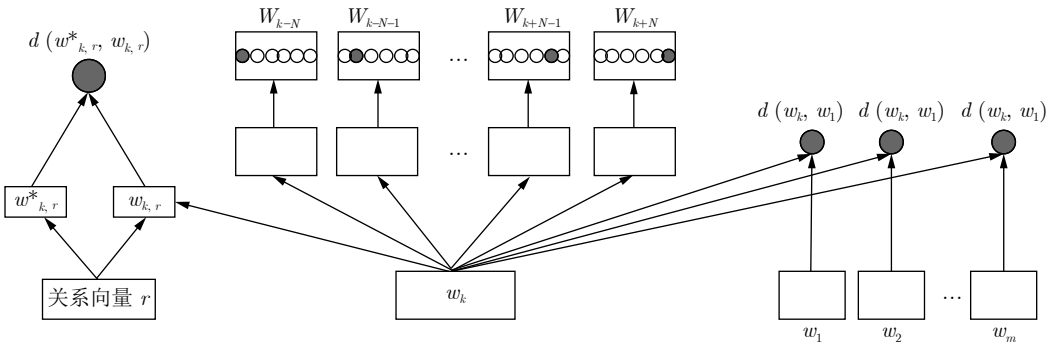


图 2　融合知识和主题信息的词向量表示模型

Fig. 2　Word vector representation model that fuses knowledge and subject information

$$r_{ij}^{z}=\sum_{j=1,j\neq i}^{n}\sum_{Y\in N_j}1,\qquad \text{如果}n_{ij}\text{与}Y\text{存在直接关系} \tag{7}$$

其中, n 为候选实体集合个数, N_j 为第 j 个候选实体集合.

2) 间接关系属性计算候选实体以三元组的形式存储在自构建的特定领域知识库中, 通过实体、关系相连接成网路状, 这种存储形式决定了候选实体间的间接关系同时存在垂直间接关系和水平间接关系. 例如在自构建的本地知识库中存在三元组: (云南, 地级市, 玉溪), (玉溪, 景点, 抚仙湖), 通过一条关系路径, 将两个三元组连接在一起, 则 "抚仙湖 "和"云南 "存在间接关系, 我们称之为垂直间接关系; 同样的, 例如本地知识库中也存在三元组: (云南, 地级市, 玉溪), (云南, 地级市, 曲靖), 如果只考虑关系路径相连接的情况, 则 "玉溪"和 "曲靖"之间并不存在关系, 这样却与事实不符. 两者之间对应同一个头实体, 也存在间接关系, 我们将这种间接关系称为水平间接关系; 同时也可以同时存在两种间接关系, 例如 (中国, 省份, 云南), (中国, 省份, 江苏), (云南, 地级市, 丽江), (丽江, 景点, 玉龙雪山), "玉龙雪山"和 "云南"存在垂直间接关系, "云南"和 "江苏"之间存在水平间接关系, 则 "玉龙雪山"和 "江苏"之间同时存在垂直和水平间接关系. 间接关系指数的计算公式为:

$$r_{ij}^{t}=\sum_{j=1,j\neq i}^{n}\sum_{Y\in N_j}\left(\frac{1}{2}\right)^{k-1}\left(\frac{1}{2}\right)^{p},\qquad \text{如果}n_{ij}\text{与}Y\text{存在直接关系} \tag{8}$$

其中, n 为候选实体集合个数, N_j 为第 j 候选实体集合, k 为路径长度, p 为水平间接次数, 例如 "玉龙雪山"和 "江苏"存在一次水平间接次数, 当两者之间存在多条路径时, 取最短路径.

1.4 特征生成模块

1.4.1 上下文特征生成

实体指称项的上下文特征可以代表指称项的文本环境, 对指称项的语义表达具有重要作用. 通过实体指称项的背景文本, 经过文本预处理 (分词、去停用词), 利用第 2.2 节训练好的融合知识和主题信息的词向量模型得到指称项的上下文向量表示. 具体方法为: 选择实体指称项所在句子经过分词、去停用词后的词作为实体指称项的上下文, 利用训练好的词表示模型得到它们的向量表示形式. 利用式 (6) 计算上下文特征相似度.

1.4.2 主题关键词特征生成

特定领域的局部特征对实体消歧具有重要作用, 例如: 在旅游领域的背景文本中, 实体指称项 "金花"的上下文信息主题围绕 "花卉名"来进行介绍, 而在文档局部上下文中主要围绕 "茶品"的金花来介绍, 可以看出局部特征对消歧有重要意义. 为了利用局部特征进行实体链接, 本文提出通过 LDA 主题模型对旅游领域背景文本的上下文进行主题分类, 利用第 2.2 节得到的融合知识和主题信息的词向量表示, 计算相同主题下的词与词之间的余弦相似度, 然后进行 K-means 聚类, 选择离聚类中心最近的 w 个词作为主题关键词, w 的取值在实验部分具体说明. 主题特征表示为:

$$S_g=\frac{\sum_{i=1}^{w}\sum_{j=1}^{z}\cos\left(E\left(w_i\right),E\left(w_j^{*}\right)\right)}{w\cdot z} \tag{9}$$

其中, $\boldsymbol{E}(w_i)$ 和 $\boldsymbol{E}(w_j^{*})$ 分别表示实体指称项主题关键词 w_i 其对应候选实体在自构建的特定领域知识库中的类别标签 w_j^{*} 的向量表示; w 为实体指称项主题关键词的个数; z 表示对应候选实体在知识库中的类别标签个数.

1.4.3 扩展词特征生成

集成实体链接相比于单实体链接充分考虑了实体之间的共现关系, 同时提高了计算效率. 利用词扩展的方法, 同时考虑 v 个实体, 充分发挥集成实体链接的优势, 具体方法为: 对于第 i 个指称项 m_i, 分别计算其他 $n-1$ 个指称项与第 i 个指称项的上下文特征和主题关键词特征的余弦相似度, 将相似度最大的 v 个实体指称项选择作为第 i 个实体指称项的扩展词, 依次迭代 n 次, 得到背景文本中每个实体指称项的扩展词. 实体指称项扩展词的计算公式为:

$$Q_k=S_w+S_g \tag{10}$$

其中, S_w 和 S_g 分别表示实体指称项的上下文相似度和主题关键词相似度; 选取 Q_k 最大的 v 个实体指称项作为本实体指称项的扩展词. v 的取值在实验部分详细说明. 扩展词特征表示为:

$$S_k=\frac{\sum_{k=1}^{v}\sum_{h=1}^{u}\cos(E(z_k),E(G_h^{*}))}{v\cdot u} \tag{11}$$

其中, $\boldsymbol{E}(z_k)$ 和 $\boldsymbol{E}(G_h^{*})$ 分别表示实体指称项扩展词和其候选实体直接三元组尾实体的向量表示; v 和 u 分别表示扩展词和其候选实体直接三元组尾实体的个数.

1.5 实体链接模块

1.5.1 关系指数计算

对于第 i 个实体指称项 m_i 和它的 v 个扩展词,

同时链接到本地特定领域知识库中的每个候选实体, 根据第 2.3.2 节的方法, 得到实体指称项候选实体与其扩展词候选实体之间的关系指数, 具体方法为: 对于候选实体 n_{ij} , 分别对它与 m_i 的 v 个扩展词的每个候选实体进行关系查找, 得到它与 v 个扩展词候选实体的关系指数之和, 最终通过归一化得到 m_i 的每个候选实体的关系指数. 计算公式表示为:

$$r_{ij} = r_{ij}^{z} + r_{ij}^{t} \tag{12}$$

依次计算出实体指称项 m_i 所有候选实体的关系指数 $r_{i1}, r_{i1}, \cdots, r_{iL}$, 其中 L 为实体指称项 m_i 的候选实体个数. 通过归一化, 得到最终的关系指数:

$$R_{ij} = \frac{r_{ij}}{r_{i1} + r_{i1}, \cdots, +r_{iL}} \tag{13}$$

1.5.2 相似度计算

相似度计算是指利用实体指称项的文本特征与知识库中候选实体的相应特征, 通过计算两者之间的余弦相似度, 以此表征实体指称项与候选实体在文本信息方面的相似度. 在本文中, 充分利用上下文相似度、主题关键词相似度和扩展词相似度, 最后得到特定领域实体指称项的相似度:

$$S_{ij} = \alpha S_{jw}^{i} + \beta S_{jg}^{i} + \gamma S_{jk}^{i} \tag{14}$$

其中, S_{jw}^{i} , S_{jg}^{i} , S_{jk}^{i} 分别表示实体指称 m_i 对其候选实体 n_{ij} 的上下文相似度、主题关键词相似度和扩展词相似度; α, β, γ 为三种相似度的权重值. 对于三种相似度权重 α, β, γ 的选择, 我们采用基于经验和权值归一的方法, 首先固定某一项特征, 观察评测效果的好坏进行其他两项权重的调整, 求出其他两项的最佳权重值, 依次固定其他特征项, 每项特征得到两个权重值, 然后将得到的每一项特征的权重值求平均并归一化得到最终的 α, β, γ . 根据实验, 我们最终设置的三项相似度权重值 α, β, γ 分别为 0.25, 0.4, 0.35. 最终将相似度与关系指数融合:

$$W_{ij} = \frac{1}{2}(R_{ij} + S_{ij}) \tag{15}$$

其中, R_{ij} , S_{ij} 分别表示实体指称项 m_i 与其候选实体 n_{ij} 的关系指数和特征相似度; 1/2 表示两者的权重值. 在文本中我们采用对等加权, 也可以考虑不对等加权的情况, 但通过初步实验结果并参考文献 [11] 表明, 少量的权值修正对实体链接结果的影响不大, 因此本文采用 1/2 作为两者的权重值.

2 实验

2.1 数据集

本文选择 Google 的开源工具包 word2vec, 采用 Skip-gram 模型作为词向量训练的基本模型, 通过提取维基百科旅游、文化分类下的文本信息, 并结合从旅游网站和百度百科、民族文化网站、中国中药杂志、中国中药材网爬取旅游信息文本 136 749 篇, 中国少数民族信息文本 95 483 篇, 药材信息文本 114 673 篇作为词表示模型的训练语料. TransE 模型的预训练使用本地特定领域知识库中的 163 759 组三元组为语料. 实验所用的测试集是本文从爬取的旅游、少数民族文化、中药材三种领域中随机分别选取 861 篇作为测试文本, 然后分别从三种领域的测试文本中人工选取含有实体歧义的文本 300 篇构建成旅游领域测试集、少数民族文化测试集和中药材测试集, 并且在每一篇文本中人工标记出领域实体指称和其在自构建的领域知识库中的对应实体, 在三个领域测试集中分别标注实体指称 1 135 个、947 个和 1 092 个, 其中旅游领域测试集、少数民族文化测和中药材测试集在自构建的领域知识库中存在对应实体对象的分别有 967 个、703 个、939 个实体指称.

2.2 实验设置与评价指标

实验的过程包括融合知识和主题信息的词向量模型训练、候选实体的生成、扩展词的生成、关系指数计算、相似度计算、实体链接等过程. 使用 jieba 分词工具实现语料预处理; 针对融合知识和主题信息的词向量模型训练, 采用 Skip-gram 模型作为词向量训练的基本模型, 窗口大小设置为 5, 设置主题词距离权重 $\alpha = 0.8$, 主题词 $m = 6$, 对于 TransE 模型的预训练, 边缘超参数设为 1, 学习速率设为 0.0001, 选用 L2 作为距离计算公式, 向量维数统一设为 200; 本文采用准确率 $P(\%)$、召回率 $R(\%)$ 和 F1 值来评估本文提出的方法, 其中文本中的实体指称项在本地知识库中存在对应实体的集合为 A; 算法输出的链接到本地知识库中实体对象上的实体指称项集合为 B. 则准确率 $P(\%)$、召回率 $R(\%)$ 和 F1 值的计算公式如下所示:

$$P(\%) = \frac{|A| \bigcap |B|}{|B|} \tag{16}$$

$$R(\%) = \frac{|A| \bigcap |B|}{|A|} \tag{17}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{18}$$

2.3 实验及其结果分析

2.3.1 实验设计

为了验证本文提出方法的可行性, 本文设置以

下 6 组实验: 实验 1: 不同相似度特征组合的实验对比. 实验 2: 验证扩展词的数量 v 对实体链接结果的影响. 实验 3: 验证主题关键词个数 w 对于实体链接准确率的影响. 实验 4: 验证不同关系属性对实体链接结果的影响. 实验 5: 本文提出的方法与目前主流的实体链接方法进行对比. 实验 6: 验证本文提出的方法在不同领域中的普适性.

2.3.2 实验结果与分析

1) 实验 1: 为了验证不同特征对实体链接结果的影响, 本实验使用旅游领域测试集, 通过选取不同的特征组合进行对比实验, 表 1 所示为不同特征组合对实验结果的影响.

表 1 不同特征组合实验结果统计

Table 1 Statistics of experimental results of different feature combinations

特征组合	P(%)	R(%)	F1
上下文	64.8	61.7	63.2
上下文+主题关键词	79.3	80.9	80.1
上下文+主题关键词+扩展词	87.7	86.5	87.1
上下文+主题关键词+扩展词+关系指数	92.6	90.4	91.5

在进行特征组合对比实验时, 使用旅游领域测试集, 主题关键词个数 $w = 4$, 扩展词个数 $v = 3$. 根据实验结果发现, 只利用上下文相似度特征和主题关键词相似度特征, 其准确率明显低于结合扩展词相似度特征和关系指数, F1 值相较于只利用上下文特征和主题特征也有明显提升, 能够达到 91.5. 分析原因主要是上下文相似度特征和主题关键词特征仅仅是基于一个实体指称项信息出发, 没有考虑一篇文章中实体指称项之间的共现信息, 并且忽略了候选实体之间的关系属性. 结合扩展词相似度特征和关系指数, 在考虑单个实体指称项的同时也充分考虑了实体指称项的共现信息和候选实体之间的关系属性, 因此准确率有了很大的提高.

2) 实验 2: 本实验在旅游领域测试集上, 分别测试扩展词个数 v 在取 1, 2, 3, 4 时对实体链接准确率的影响, 实验结果如表 2 所示.

在进行扩展词个数实验时, 使用旅游领域测试集, 同时考虑上下文特征、主题关键词特征、扩展词特征、关系指数, 主题关键词个数 $w = 4$. 根据实验结果发现, 扩展词的个数对实体链接结果有较大影响, F1 值可以从最低的 83.1 提升到 91.5, 并且相比于只利用上下文特征和主题关键词特征的 F1 值, 有了较大提升, 说明加入扩展词特征可以对实体链接有较大帮助. 从实验结果表明, 当扩展词个数 $v = 3$ 时, F1 值达到最大值 91.5.当个数大于 3 时准确率和 F1 值都有所降低. 分析原因主要是因为当扩展词个数太小时, 不仅没有充分利用实体指称项之间的共现信息, 并且会影响候选实体的关系指数, 所以准确率会降低, 当扩展词个数太大, 会出现冗余信息, 对实体指称项的信息表达和候选实体关系指数计算都会产生不好的影响. 所以本文扩展词个数取 $v = 3$.

表 2 不同 v 值实验结果统计

Table 2 Statistical results of different v values

v	P(%)	R(%)	F1
1	84.4	81.8	83.1
2	89.6	87.1	88.3
3	92.6	90.4	91.5
4	90.3	89.4	89.8

3) 实验 3: 本实验在旅游领域测试集上, 分别测试主题关键词个数 w 在取 1, 2, 3, 4, 5 时对实体链接准确率的影响, 实验结果如表 3 所示.

表 3 不同 w 值实验结果统计

Table 3 Statistical results of different w values

w	P(%)	R(%)	F1
1	86.4	83.5	84.9
2	89.8	87.6	88.7
3	90.5	89.2	89.8
4	92.6	90.4	91.5
5	89.7	88.6	89.1

在进行主题关键词个数实验时, 使用旅游领域测试集, 同时考虑上下文特征、主题关键词特征、扩展词特征、关系指数, 扩展词个数 $v = 3$. 通过对比不同主题词个数 w 和不同扩展词个数 v 的对比实验表明, 扩展词特征与主题词特征的作用基本相当, 最小 F1 指分别为 83.1 和 84.9, 但是主题词不同个数之间 F1 值的差距没有不同扩展词个数之间明显. 根据实验结果发现, 当主题关键词个数 $w = 4$ 时, F1 值达到最大值 91.5, 当个数大于 4 时准确率降低. 分析原因在于提取主题关键词时采用聚类的方法, 当主题关键词个数太小时, 无法代表领域特定信息, 当个数大于 4 时, 又造成信息冗余, 将多余信息引入到相似度计算中, 从而导致实体链接的 F1 值下降. 所以本文主题关键词个数取 $w = 4$.

4) 实验 4: 为了验证关系属性中每个子属性的效果对实体链接结果的影响, 本实验使用旅游领域测试集, 通过依次增加关系属性中各个子属性来设置对比实验, 观察实验结果如表 4 所示.

在进行各关系子属性的实验时, 使用旅游领域

表 4　各个关系子属性的实验结果统计
Table 4　Statistical results of experimental results for each relationship sub-attribute

关系属性	P(%)	R(%)	F1
直接关系	89.3	87.2	88.2
直接关系+垂直间接关系	91.8	88.7	90.2
直接关系+水平间接关系	91.1	87.6	89.3
直接关系+两个间接关系	92.6	90.4	91.5

测试集, 同时考虑上下文特征、主题关键词特征、扩展词特征, 扩展词个数 $v=3$, 主题词个数 $w=4$. 实验结果表明, 利用候选实体之间的直接关系使得实体链接的 F1 值有了较小提升, 分析原因是自构建的特定领域知识库中并不完整, 只利用直接关系信息对实验结果帮助有限, 同时通过水平间接关系和垂直间接关系的实验结果对比, 垂直间接关系对实体链接结果影响更大, 说明通过关系路径相连的候选实体之间的关系信息对实体链接更有帮助, 但是通过最终的实验结果表明, 将两种间接关系同时考虑, 更能增加候选实体的关系信息, 对实体链接帮助更大.

5) 实验 5: 为了验证本文提出方法的可行性, 在旅游领域测试集上, 将本文的方法与其他几种实体链接方法进行比较, 实验结果如表 5 所示.

表 5　本文方法与其他方法的比较
Table 5　Comparison of methods in this paper with other methods

方法名	P(%)	R(%)	F1
Wikify	71.3	73.9	72.6
Cucerzan	76.5	80.2	78.3
SVM[27]	83.1	85.3	84.2
Score[28]	87.4	86.5	86.9
EAT[21]	80.7	82.9	81.8
Zero-shot[29]	91.4	88.0	89.7
本文的方法	92.6	90.4	91.5

在旅游领域测试集中将以上基线方法复现, 其中参数设置与其论文中相同. 根据实验结果表明, 本文提出的方法与传统的统计机器学习的方法相比较 F1 值有明显的提升, 并且不需要标注语料, 更简洁高效; 与 EAT[21] 方法相比较, Moreno 等[21] 通过扩充知识库中实体的锚文本对文本中的单词和知识库中的实体在同一个向量空间中学习指称项与候选实体的向量表示, 并通过训练分类器进行实体链接, 两种方法都是基于词嵌入, 本文的方法准确率有较大提升, 我们分析原因在于我们的语料主要是针对特定领域, 语料数据集规模相较于公共数据集偏小, 所以词嵌入效果没有达到最佳, 但是我们的方法在词嵌入的基础上, 将知识和主题信息融入词向量表示中, 将文本信息和知识库信息融合, 同时综合考虑了上下文特征、主题特征、词扩展特征、关系指数特征, 所以比 EAT[21] 方法在 F1 值上有了较大的提高, 也验证了本文的方法更适应于语料偏少的特定领域; 与 Zero-shot[29] 相比较, 前者利用的是最新的神经网络模型, 与它相比较 F1 值有较小提高, 证明了本方法达到了较高水平, 也证明了本方法在对特定领域实体链接任务的可行性.

6) 实验 6: 为了验证本文提出的方法在不同领域中的普适性, 将本文的方法在旅游领域测试集、少数民族文化测试集和中药材测试集中进行比较, 实验结果如表 6 所示.

表 6　不同领域的实验结果统计
Table 6　Statistics of experimental results in different fields

领域名称	P(%)	R(%)	F1
旅游领域	92.6	90.4	91.5
少数民族领域	91.8	89.6	90.7
药材领域	90.3	91.4	90.8

由实验结果表明, 在不同的领域语料中的 F1 值变化不大, 其中在旅游领域中的 F1 值最大, 在少数民族和药材领域 F1 值基本一致, 分析原因: 在旅游领域中, 由于其关系类别少、实体个数多的特点, 其扩展词可以很好地表征其语义信息, 利用扩展与实体指称项的候选实体之间的关系信息也比较明显. 但是在少数民族和药材领域, 关系种类更加复杂, 实体与实体之间的关系信息也不明显, 所以在这两种领域中, 扩展词特征和关系指数不如在领域领域中明显, 造成了 F1 值略有下降. 但是从不同领域的对比实验中表明, 本文方法针对标注语料少, 流行度等消歧特征不明显的问题, 在不同特定领域中的效果基本稳定并且有较好的 F1 值.

3　总结和展望

本文针对现有的实体链接方法无法将文本信息和本地知识库信息充分相结合, 提出了一种简单高效的基于关系指数和表示学习的特定领域集成实体链接方法. 利用表示学习将文本信息和知识库信息相融合, 简单高效且适应于特定领域语料偏少的特点. 实验结果表明, 该方法与现有的实体链接方法相比, 不需要标注语料, 其实体链接准确率和 F1 值比较理想, 同时更适应于语料偏少的特定领域. 下一步的工作是对已经构建的小规模特定领域知识库

进行扩充和完善, 同时不断挖掘领域文本中特有的属性特征, 改进实验效果.

References

1 Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. 782−792

2 Burdick D, Fagin R, Kolaitis P G, et al. Expressive power of entity-linking frameworks. *Journal of Computer and System Sciences*, 2019, **100**: 44−69

3 Saeedi A, Peukert E, Rahm E. Using link features for entity clustering in knowledge graphs. In: Proceedings of European Semantic Web Conference. Springer, Cham, 2018. 576−592

4 Dubey M, Banerjee D, Chaudhuri D, et al. Earl: Joint entity and relation linking for question answering over knowledge graphs. In: Proceedings of International Semantic Web Conference. Springer, Cham, 2018. 108−126

5 Bunescu R C. Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of Conference of the European Chapter of the Association for Computational Linguistics, 2006. 9−16

6 Ganea O E, Ganea M, Lucchi A, et al. Probabilistic bag-of-hyperlinks model for entity linking. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016. 927−938

7 Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007. 708−716

8 Nguyen H T, Cao T H. Exploring wikipedia and text features for named entity disambiguation. In: Proceedings of Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2010. 11−20

9 Zeng Y, Wang D, Zhang T, et al. Linking entities in short texts based on a Chinese semantic knowledge base. *Natural Language Processing and Chinese Computing*. Springer, Berlin, Heidelberg, 2013. 266−276

10 Han X, Sun L, Zhao J. Collective entity linking in web text: A graph-based method. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011. 765−774

11 Liu Q, Zhong Y, Li Y, et al. Graph-based collective Chinese entity linking algorithm. *Comput. Res. Develop*, 2016, **53**(2): 270−283

12 Ferragina P, Scaiella U. Tagme: On-the-fly annotation of short text fragments. In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010. 1625−1628

13 Francis-Landau M, Durrett G, Klein D. Capturing semantic similarity for entity linking with convolutional neural networks. In: Proceedings of the North American Chapter of the Association for Computational Linguistics, 2016. 1256−1261

14 Wang W, Arora R, Livescu K, et al. On deep multi-view representation learning: Objectives and optimization. In: Proceedings of International Conference on Machine Learning 2015, 2016. 1083−1092

15 Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013, **35**(8): 1789−1828

16 Zeng Qi, Zhou Gang, Lan Ming-Jing, et al. Polysemous word multi-embedding calculation. *Journal of Chinese Computer Systems*, 2016, **37**(5): 1417−1421

17 Raiman J R, Raiman O M. DeepType: Multilingual entity linking by neural type system evolution. In: Proceedings of 32nd AAAI Conference on Artificial Intelligence, 2018. 5406-5413

18 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations, 2013. 1−12

19 Goldberg Y, Levy O. Word2vec Explained: Deriving Mikolov et al.′ s negative-sampling word-embedding method. CoRR abs/1402.3722, 2014, 1−5

20 Kar R, Reddy S, Bhattacharya S, et al. Task-specific representation learning for web-scale entity disambiguation. In: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence, 2018. 5812−5819

21 Moreno J G, Besancon R, Beaumont R, et al. Combining Word and Entity Embeddings for Entity Linking. In: Proceedings of European Semantic Web Conference. Springer, Cham, 2017. 337−352

22 You M, Yang H, Lin Z, et al. BTM topic modeling approach to named entity linking. *Journal of Physics: Conference Series. IOP Publishing*, 2018, **1060**(1): 012−027

23 Gao Y, Li A, Duan L. Entity disambiguation method based on multi-feature fusion graph model for entity linking. *Application Research of Computers*, 2017: 2909−2914

24 Sakor A, Mulang I O, Singh K, et al. Old is gold: Linguistic driven approach for entity and relation linking of short text. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 2336−2346

25 Xu C, Bai Y, Bian J, et al. Rc-net: A general framework for incorporating knowledge into word representations. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. ACM, 2014. 1219−1228

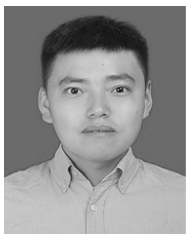
26 Ganea O E, Ganea M, Lucchi A, et al. Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 2787−2795

27 Bao-Xing H, Bao T F, Zhu H S, et al. Topic modeling approach to named entity linking. *Journal of Software*, 2014, (9): 2076−2087

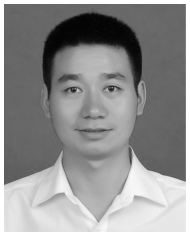
28 Wu Y B, Zhu D H, Liao X W, et al. Knowledge graph reasoning based on paths of tensor factorization. *Pattern Recognition and Artificial Intelligence*, 2017, **30**(5): 473−480

29 Kundu G, Sil A, Florian R, et al. Neural cross-lingual corefer-

ence resolution and its application to entity linking. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 395−400



蒋胜臣 昆明理工大学信息工程与自动化学院硕士研究生. 主要研究方向为自然语言处理, 知识图谱.
E-mail: jsc_study@hotmail.com
(**JIANG Sheng-Chen** Master student at Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing and knowledge graph.)



王红斌 博士, 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为智能信息系统, 自然语言处理, 信息检索.
E-mail: whbin2007@126.com
(**WANG Hong-Bin** Ph.D., associate professor at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers intelligent information system, natural language processing, information retrieval.)

余正涛 博士, 昆明理工大学信息工程与自动化学院教授. 主要研究方向为自然语言处理, 机器翻译, 信息检索.本文通信作者.
E-mail: ztyu@hotmail.com
(**YU Zheng-Tao** Ph.D., professor at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing, machine translation, and information retrieval. Corresponding author of this paper.)

线岩团 昆明理工大学信息工程与自动化学院博士研究生. 主要研究方向为自然语言处理, 信息抽取, 机器翻译.
E-mail: yantuan.xian@gmail.com
(**XIAN Yan-Tuan** Ph.D. candidate at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing, information extraction, and machine translation.)

王红涛 昆明理工大学信息工程与自动化学院硕士研究生. 主要研究方向为自然语言处理, 信息抽取.
E-mail: 15893739522@163.com
(**WANG Hong-Tao** Master student at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers natural language processing, and information extraction.)