计算机科学
COMPUTER SCIENCE

基于自注意力模型的本体对齐方法

吴子仪，李邵梅，姜梦函，张建朋

引用本文

吴子仪，李邵梅，姜梦函，张建朋. 基于自注意力模型的本体对齐方法[J]. 计算机科学, 2022, 49(9): 215-220.

WU Zi-yi, LI Shao-mei, JIANG Meng-han, ZHANG Jian-peng. Ontology Alignment Method Based on Self-attention[J]. Computer Science, 2022, 49(9): 215-220.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210700190

基于自注意力模型的本体对齐方法

吴子仪　李邵梅　姜梦函　张建朋
国家数字交换系统工程技术研究中心　郑州 450002
(mushaboommm@163.com)

摘　要　随着知识图谱在人工智能领域的发展,对不同源的知识图谱进行融合,以得到覆盖范围更广的知识图谱的需求日益增加。本体作为知识图谱的上层结构,对知识图谱的构建具有指导作用。为了解决知识图谱融合中本体对齐的问题,文中提出了基于自注意力模型融合多维相似度的方法,从而提高本体对齐的精度。首先,对来自两个本体的概念进行基于字符串的、基于语义的和基于结构信息的多维度相似性度量;然后,使用自注意力模型对上述多种相似度度量结果进行融合,进而判断是否相似并进行对齐。在公开数据集上进行实验,实验结果表明,相比现有的本体对齐方法,所提方法通过聚合多维度的相似性特征能够得到更优的对齐结果。

关键词:知识图谱融合;本体对齐;相似度计算;自注意力模型
中图法分类号　TP391.1

Ontology Alignment Method Based on Self-attention

WU Zi-yi,LI Shao-mei,JIANG Meng-han and ZHANG Jian-peng
National Digital Switching System Engineering & Technological R & D Center,Zhengzhou 450002,China

Abstract　With the development of knowledge graph in the field of artificial intelligence,there is an increasing demand to integrate knowledge graph from different sources to obtain a big knowledge graph with wider coverage. Ontology is the superstructure that can guide the construction of knowledge graph. To solve the problem of ontology alignment in knowledge graph fusion,this paper proposes an ontology alignment method based on self-attention model to combine multidimensional similarities. Firstly,two concepts from two ontologies are multi-dimensional measured by string-based,semantic-based and structure-based similarities. Then,self-attention model is used to combine above similarity calculations to judge whether the two concepts are similar or not and align them. Experiments on public datasets show that,compared with existing ontology alignment methods,the proposed method can obtain better alignment results by aggregating multi-dimensional similarity features.

Keywords　Knowledge graph fusion,Ontology alignment,Similarity calculation,Self-attention model

1　概述

知识图谱是一种语义网络,旨在从数据中识别、发现和推断事物、概念之间的复杂关系,是事物关系的可计算模型。随着以知识图谱为支撑的智能系统的快速发展,知识图谱在人工智能中的重要性日益凸显,被看作是人工智能的基石。因此,如何对多种来源的知识图谱进行融合,得到覆盖率更广的知识图谱成为了一个重要的研究方向。本体作为知识图谱的上层结构,是共享概念的形式化、概念化描述。本体对齐作为知识图谱融合中的关键技术,受到了广大研究人员的关注。

本体对齐指在不同本体中,在具有相关语义的概念之间建立对应关系的过程[1],以解决不同本体间的知识共享,并提高相互操作性[2]。本体对齐技术通常从不同本体间概念的相似度出发进行研究。对本体中概念相似度的计算又划分为元素级和结构级,其中元素级只利用概念自身的信息,而未利用概念之间的关系;结构级则利用本体层次结构的信息。基于元素级的本体对齐方面,Xu 等[3]使用词典中心词、近义词、距离计算等方法进行相似度的计算,Yao 等[4]提出了结合实例层的相似度计算;基于结构级的本体对齐方面,Yu 等[5]利用哈斯图来获取语义信息;Jiang 等[6]通过聚类和模块化,将本体划分为子本体再使用信息检索策略进行对齐。

考虑到单一维度的度量不能准确刻画两个本体中概念相似度的问题,Euzenat 等[7]首次利用多种度量方法来计算概念之间的相似度,然后将这些相似度进行加权,通过加权后的相似度值来判断并对本体进行对齐。后续围绕如何更有效地对多种不同的相似度度量结果进行融合的问题,研究人员提出了基于机器学习的本体对齐方法[8],这些方法把本体对齐看成是分类问题,将概念间的多种相似性度量值作为特征,利用

到稿日期:2021-07-19　返修日期:2022-02-28
基金项目:国家自然科学基金青年科学基金(62002384);郑州市协同创新重大专项(162/32410218)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(62002384) and Zhengzhou Collaborative Innovation Major Project(162/32410218).
通信作者:李邵梅(13513127249@163.com)

分类模型来实现对两个概念是否相似的判决。Alboukaey等[9]计算了两两概念间多种不同元素级的相似度，并将其作为特征，然后使用回归模型进行分类；Lev等[10]计算了更多的相似性度量，使用逻辑回归、随机森林和极致梯度等机器学习算法进行分类；Sengodan等[11]使用基于阈值的支持向量机和基于语义增强最邻近算法进行分类。

已有研究表明，两个概念间多种相似性度量融合的方法能有效提高本体对齐效果。为此，围绕如何提取更多维度的相似性度量以及如何对这些度量结果进行有效融合的问题，本文提出了基于自注意力模型对两两概念间多种元素级和结构级相似度进行融合的本体对齐方法。首先分别基于字符串、语义和结构信息计算两个本体间概念的相似度；然后将这些不同类别的相似性度量作为输入，送入自注意力模型中进行融合，根据融合的结果判断概念是否相似，进而进行本体对齐。已有的基于本体多度量特征的分类方法通常使用的是传统机器学习模型[12]，例如文献[10]使用了逻辑回归、随机森林和极致梯度等，忽略了不同相似性度量之间的相关性，对特征的挖掘不够深入，而本文使用自注意力模型能够自动学习不同的相似性度量对概念间相似性判断的重要性权重，进而更好地聚合各个相似性特征，得到更佳的本体对齐结果。

2 本体对齐问题描述

本体通常表示为 $O(C,P,H)$，C 为类集合，P 为属性集合，H 为类的层次关系[13]。类和属性都被称为概念[14]。本体对齐的目的是找到源本体和目标本体间的类及属性的对齐[15]。对于给定的两个本体 $O_1(C_1,P_1,H_1)$ 和 $O_2(C_2,P_2,H_2)$，本体对齐任务就是将两个本体中的类和属性分别构建匹配单元进行匹配。为此，定义匹配单元对为 $(c_1,c_2,cor(c_1,c_2))$ 和 $(p_1,p_2,cor(p_1,p_2))$ 两类，其中，$c_1\subset C_1$，$c_2\subset C_2$，$p_1\subset P_1$，$p_2\subset P_2$。假设 O_1 中有 a 个类，O_2 中有 b 个类，则 $(c_1,c_2,cor(c_1,c_2))$ 包含 $a\times b$ 个基于类的匹配单元对，属性同理。$cor(c_1,c_2)$，$cor(p_1,p_2)$ 是对齐关系的置信度，超过置信度阈值（该阈值由分类模型学习所得）的匹配单元对被认为是对齐的。本文只研究一对一的等价关系，每一组匹配单元对都被分配一个标签“0”或“1”，其中，“0”表示匹配单元对不对齐，“1”表示匹配单元对对齐。进而，本体对齐被简化为一个二分类问题。

3 多维度的概念间相似性度量

如前文所述，本体对齐的主要任务就是计算两个本体间两两概念的相似度并进行判决，即对每个类匹配单元对和属性匹配单元对进行相似性度量和判决。为了对匹配单元对的相似性进行全面、准确的度量，本文从字符级、语义级、结构级等不同维度出发，分别计算匹配单元对中两两概念的对齐关系置信度，并将其作为后续进行相似性融合判决的基础。

3.1 基于字符串的相似性度量

基于字符串的相似性度量是操作于字符串序列或字符组合，以字符串共现和重复程度为相似度的衡量标准[16]。

对于匹配单元对中的两个概念，类 c_1 和 c_2 或者属性 p_1 和 p_2，分别采用如下指标计算该匹配单元对的类与类之间或属性与属性之间的基于字符串的相似性度量：N-gram 1、N-gram 2、N-gram 3、N-gram 4、Dice 系数（Dice Coefficient）、Jaccard 相似性（Jaccard Similarity）、Jaro 距离（Jaro Measure）、最长公共子串（Longest Common Substring）、Monge-Elkan、Smith-Waterman、Needleman-Wunsh、Affine 间隙（Affine Gap）、Bag 距离（Bag Distance）、余弦相似性（Cosine Similarity）、部分比率（Partial Ratio）、软 TF-IDF（Soft TF-IDF）、Editex、广义 Jaccard 相似性（GeneralizedJaccard）、Jaro-Winkler、Levenshtein 距离（Levenshtein Distance）、部分标记排序（Partial Token Sort）、模糊比率（Fuzzy Wuzzy Ratio）、Soundex、TF-IDF、标记排序（Token Sort）、Tversky 指数（Token Sort）、重叠系数（Longest Common Subsequence）[10]。

3.2 基于语义的相似性度量

由于不同团队开发的本体之间可能存在同义词[17]，如“trade”和“business”，而上述基于字符串的相似性度量通常难以描述这种深层次的对应关系，因此在本体对齐过程中，还需要挖掘概念间语义层面的相似性[18]。

对于匹配单元对中的两个类 c_1 和 c_2 或者属性 p_1 和 p_2，分别使用 Wordnet[19] 和 Word2vec[20] 对类或属性的文本内容进行基于语义的相似性度量。Wordnet 是以同义词集合为基础构建的词库，能反映词与词之间存在的关系。其中，对于类或属性由多个单词组成的情况，取两个类或属性的所有可能的单词集对的最大相似度值。Word2vec 则是将词投射到向量空间，语义相近的词语在向量空间中的距离也相近。其中，对于类或属性由多个单词组成的情况，使用 Word2vec 变体的 Doc2vec[21] 来提取类或属性的语义表征。

3.3 基于结构的相似性度量

本体中除了类或属性的文本内容，结构信息也可以用于本体对齐。对齐的类或属性通常具有相似的结构[14]，如具有相似的父类或在层次结构中的相似位置。

为此，本文在对齐过程中也考虑了两两概念间结构的相似性。对于匹配单元对中的两个类 c_1 和 c_2 或者属性 p_1 和 p_2，分别从其所在的本体中检索这两个类或属性的父类和路径。其中，对于类来说，类的父类是与“supClassOf”相连的类，类的路径是从初始类到当前类的能描述该类整个层次结构的字符串。例如，图 1 给出了可视化本体的一部分，对于类“Monograph”，其父类为“Book”，其路径为“Thing/Reference/Book/Monograph”。

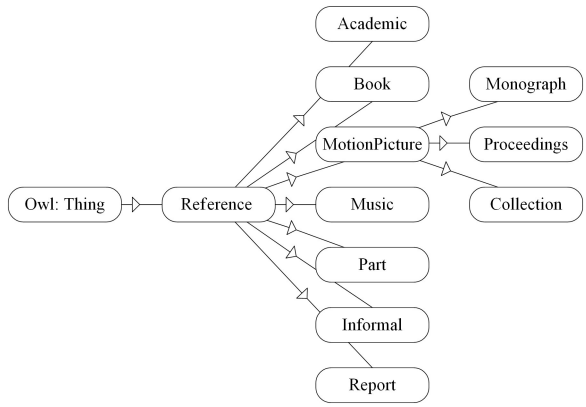


图 1 类的父类和路径

Fig. 1 Parent class and path of a class

对于属性来说，属性的父类就是该属性所描述的类的

名称，属性的路径就是其所描述的类的路径。例如，图 2 给出了本体文件中描述属性“chapters”的片段，可以看出其描述的类为“Reference”，即“Reference”为该属性的父类，则其路径为“Thing/Reference/Book”。

```
〈owl:ObjectProperty rdf:ID="chapter"〉
〈rdfs:domain rdf:resource="#Reference"/〉
〈rdfs:range rdf:resource="#Chapter"/〉
〈rdfs:label xml:lang="en"〉chapters〈/rdfs:label〉
〈rdfs:comment xml:lang="en"〉The chapters of a book(monograph or collection).〈/rdfs:comment〉
〈/owl:ObjectProperty〉
```

图 2 本体文件中描述属性的片段

Fig. 2 Fragments describe properties in ontology file

假设匹配单元对中 c_1 或 p_1 的父类和路径分别是 f_1 和 r_1，c_2 或 p_2 的父类和路径分别是 f_2 和 r_2。那么分别对两个概念的父类 f_1 和 f_2 按照 3.1 节和 3.2 节中的方法进行基于字符串和基于语义的相似性计算；分别对两个概念的路径 r_1 和 r_2 按照 3.1 节和 3.2 节中的方法进行基于字符串和基于语义的相似性计算。然后，把父类和路径得到的相似性度量结果拼接，得到最终的两个概念间的结构相似性度量。

3.4 两个本体间的相似性特征矩阵

假定本体 O_1 和本体 O_2 间共有 N 个匹配单元对，对于每个匹配单元对 i，首先使用 3.1－3.3 节中的方法分别进行相似性计算，得到 87 维的相似性特征向量 $\boldsymbol{X}_i$：

$$\boldsymbol{X}_i=[sim1_i,sim2_i,sim3_i],i=1,\cdots,N \tag{1}$$

$$\boldsymbol{X}_i=[sim1_i,sim2_i,sim3_i],i=1,\cdots,N$$

其中，$\boldsymbol{sim}1_i=[sim1_i^1,\cdots,sim1_i^{N_1}]$是利用 3.1 节中的方法计算得到的匹配单元对中的两个类或者属性间的字符串相似性特征向量，如表 1 中的“基于字符串的相似性度量”所示，共 27 维；$\boldsymbol{sim}2_i=[sim2_i^1,\cdots,sim2_i^{N_2}]$是利用 3.2 节中的方法计算得到的匹配单元对中的两个类或者属性间的文本内容相似性特征向量，如表 1 中的“基于语义的相似性度量”所示，共 2 维；$\boldsymbol{sim}3_i=[sim3_i^1,\cdots,sim3_i^{N_3}]$是利用 3.2 节中的方法计算得到的匹配单元对中的两个类或者属性间的结构相似性特征向量，如表 1 中的“基于结构的相似性度量”所示，共 58 维。

表 1 相似性度量

Table 1 Similarity measures

相似性度量类型	使用的相似性度量指标	数量
基于字符串的相似性度量	N-gram 1，N-gram 2，N-gram 3，N-gram 4，Dice 系数，Jaccard 相似性，Jaro 距离，最长公共子串，Monge-Elkan，Smith-Waterman，Needleman-Wunsh，Affine 间隙，Bag 距离，余弦相似性，部分比率，软 TF-IDF，Editex，广义 Jaccard 相似性，Jaro-Winkler 距离，Levenshtein 距离，部分标记排序，模糊比率，Soundex，TF-IDF，标记排序，Tversky 指数，重叠系数	27
基于语义的相似性度量	Wordnet 相似性，Word2vec/Doc2vec 相似性	2
基于结构的相似性度量	概念的父类之间所有基于字符串和基于语义的相似性度量，概念路径之间所有基于字符串和基于语义的相似性度量	58

然后，综合 N 个匹配单元对的相似性特征向量构建本体 O_1 和本体 O_2 的相似性特征矩阵 $\boldsymbol{X}$。

通过上述步骤，我们可以得到本体 O_1 和本体 O_2 间的相似性特征矩阵，基于该矩阵，后续将利用自注意力模型进行本体对齐，具体内容如第 4 节所述。

4 基于自注意力模型的本体对齐

为了有效融合第 3 节中描述的两个本体间两两概念的多维度相似性度量进行本体对齐，本文采用自注意力模型[22]自动学习每种相似性度量对最终对齐结果的贡献，完整流程如图 3 所示。

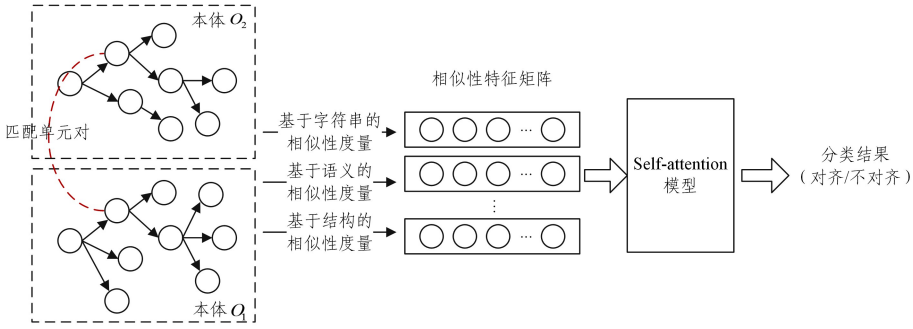


图 3 基于自注意力模型的本体对齐流程

Fig. 3 Ontology alignment process based on self-attention model

4.1 自注意力模型的原理

对于输入信息 $\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]\in R^{D_x\times N}$，其中 $n\in[1,N]$ 表示一组输入信息。对于每个输入序列，通过与 3 个不同的权值矩阵 $\boldsymbol{W}^Q$，$\boldsymbol{W}^K$ 和 $\boldsymbol{W}^V$ 相乘，将其映射到 3 个不同的向量空间中，分别得到查询向量(Query)$\boldsymbol{Q}$、键向量(Key)$\boldsymbol{K}$ 和值向量(Value)$\boldsymbol{V}$。

$$\boldsymbol{Q}=\boldsymbol{W}^Q\boldsymbol{X}\in R^{D_k\times N} \tag{2}$$

$$\boldsymbol{K}=\boldsymbol{W}^K\boldsymbol{X}\in R^{D_k\times N} \tag{3}$$

$$\boldsymbol{V}=\boldsymbol{W}^V\boldsymbol{X}\in R^{D_v\times N} \tag{4}$$

其中，D_k 为 $\boldsymbol{K}$ 的维度，D_v 为 $\boldsymbol{V}$ 的维度。

使用缩放点积作为注意力打分函数，得到注意力输出向量序列为：

$$\boldsymbol{H}=\boldsymbol{V}\,\mathrm{softmax}\left(\frac{\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}}{\sqrt{D_k}}\right) \tag{5}$$

其中，$\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}$ 是计算注意力权重的过程，$\boldsymbol{H}=[\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n]$。

最后，对于注意力层的输出 $\boldsymbol{h}_n$，连接一个全连接层进行二分类。

在本文的应用中，上述 $\boldsymbol{X}$ 就是 3.4 节中描述的两个本体间

的相似性特征矩阵。

4.2 Focal loss 损失函数

由于匹配单元对的标签反映的是本体 O_1 和本体 O_2 的两两类间或者两两属性间的相似度，通常本体间可对齐的单元对占少数，因此导致标签为“1”的匹配单元对数量少于标签为“0”的匹配单元对数量。为此，在采用上述自注意力模型对其进行分类时就存在数据类别不均衡的问题。

为了解决这个问题，本文在使用自注意力模型进行训练和测试的过程中使用了文献[23]提出的一种改进的交叉熵损失函数 Focal loss，其具体公式如式(7)所示，其中，y 是真实类别标签，y' 是预测样本类别为 1 的概率，$y' \in [0,1]$。Focal loss 在原有的交叉熵损失函数中加入了因子 γ 和 α，其中 γ 让易分样本的损失变小，从而使得模型更加关注难分样本，α 用于平衡正负样本比例。

$$L_{fl}=\begin{cases}-\alpha(1-y')^{\gamma}\log y', & y=1\\ -(1-\alpha)y'^{\gamma}\log(1-y'), & y=0\end{cases} \tag{6}$$

5 实验与评估

5.1 实验数据集

本文使用了来自本体对齐竞赛 OAEI(Ontology Alignment Evaluation Initiative)的两个标准测试集进行实验，数据集的具体参数如表 2 所列。数据集 1 是来自 Benchmarks 的关于书目引用的本体集合，包含了 7 个本体，#101 是参考本体，其他本体(#102，#103，#301，#302，#303，#304)是与 #101 进行对齐比较的本体，该数据集还包含了 6 个对齐(#101－#102，#101－#103 等)文件，一些匹配单元对间对齐的具体例子如表 3 所列，其中，3 个对齐文件用于训练，3 个对齐文件用于测试。数据集 2 包括来自 Benchmarks 的几个本体和关于会议组织的所有本体集合，包含了 27 个本体和 26 个对齐文件，其中，8 个对齐文件用于训练，18 个对齐文件用于测试(不同对齐文件包含的匹配单元对数不同，实际用于训练的匹配单元数和用于测试的匹配单元对数比例为 1.2∶1)。

表 2 数据集

Table 2 Datasets

分组	本体数	对齐文件数	匹配单元对数
数据集 1	7	6	29088
数据集 2	27	26	169393

表 3 #101－#302 对齐例子

Table 3 Examples of alignment in #101－#302

#101	#302	备注
Collection	Book	类
TechReport	TechReport	类
Report	Publication	类
Reference	Resource	类
date	publishedOn	属性

5.2 评估标准

本体对齐的评价指标为 F-Measure，F-Measure 是结合精确率 p(式(9))和召回率(式(10))r 的综合评定指标，具体计算式如式(8)所示：

$$F=2\cdot\frac{p\cdot r}{(p+r)} \tag{7}$$

$$p=\frac{|R\cap A|}{|A|} \tag{8}$$

$$r=\frac{|R\cap A|}{|R|} \tag{9}$$

其中，A 是利用对齐算法得到的正确对齐分类结果，R 是对齐文件中的真实对齐结果。

5.3 实验内容

为了验证本文方法的有效性，分别在 5.1 节介绍的数据集 1 和数据集 2 上进行实验。

5.3.1 数据预处理

为了进一步缓解 4.2 节中提到的类别不均衡问题带来的影响，在训练模型的过程中使用 SMOTEENN 算法对训练数据进行过采样。

SMOTE 算法(Synthetic Minority Oversampling Technique)是通过合成少数类样本来增加少数类样本数量的方法。其合成策略为：对于每个少数类样本 a，随机选择一个 a 的最邻近样本 b，在 a 和 b 的连线上随机找到一点作为新的少数类样本。但该种算法容易放大原数据集的噪声样本，从而导致过拟合的出现。SMOTEENN 算法是 SMOTE 算法的改进，它在进行 SMOTE 操作之后，又对样本进行了 ENN 清洗，其清洗规则为：对于多数类样本 c，如果 c 的 K 个近邻点有超过一半都不属于多数类，那么就去除该样本。

5.3.2 实验参数设置

本文实验的模型参数设置分别为：训练集样本数量 $batch\_size=128$，训练轮数 $num\_epochs=10$，每个匹配单元对的相似性特征向量维度 $feature\_size=87$，注意力头数 $num\_head=1$，自注意力模型隐藏单元数 $hidden\_num=1\,024$，学习率 $learning\_rate=5\times10^{-4}$，$dropout$ 比率=0.5。

5.3.3 对比实验

根据公开文献上可以找到的在数据集 1 上进行本体对齐的结果，将本文方法与文献[10,24-25]中的方法进行了对比实验，文献[10]分别使用了基于逻辑回归算法、随机森林算法、极致梯度提升算法进行分类，分别表示为“LR”“RF”“XGBoost”；文献[24]使用了基于关联规则的本体对齐方法，表示为“FOAM”；文献[25]使用了基于决策树的分类方法，表示为“DT”。

如表 4 所列，由实验结果可以看出，相比其他方法，本文提出的基于自注意力模型的本体对齐方法在 3 个本体对上的对齐性能均有所提升，平均的 F-Measure 值达到了 0.96。这是因为：一方面，相比“FOAM”和“DT”，本文使用了更多的相似性度量，利用了更为丰富的匹配单元对间的相似性特征；另一方面，相比逻辑回归、随机森林和极致梯度这类传统的机器学习算法，本文使用自注意力模型可以自动识别不同相似性度量的贡献，并对其进行有效融合，从而对匹配单元对进行更准确的对齐。

表 4 数据集 1 上使用不同本体对齐方法的 F-Measure

Table 4 F-Measure of different ontology alignment methods in dataset 1

对齐本体	本文方法	LR[10]	RF[10]	XGBoost[10]	FOAM[24]	DT[25]
#101－#302	0.98	0.72	0.71	0.72	0.77	0.759
#101－#303	0.96	0.82	0.82	0.75	0.84	0.816
#101－#304	0.95	0.90	0.91	0.91	0.95	0.960
平均	0.96	0.81	0.81	0.79	0.85	0.845

根据公开文献上可以找到的对数据集 2 进行本体对齐的结果，在数据集 2 上，将本文方法与文献[9-10]中的方法进行了比较，其中，文献[9]基于元素级相似度，使用了多层感知器算法、REP 树算法和 M5 Rules 算法作为分类模型，分别用“MP”“REPTree”和“M5”表示。

如表 5 所列，在数据集 2 上，本文方法的 F-Measure 值为 0.71，高于其他方法。但相比数据集 1，数据集上的 F-Measure 值较低，这是因为数据集 2 的规模更大、结构更加复杂。

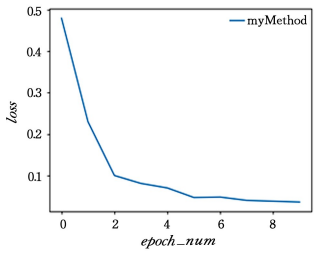
表 5 数据集 2 上使用不同本体对齐方法的 F-Measure

Table 5 F-Measure of different ontology alignment methods in dataset 2

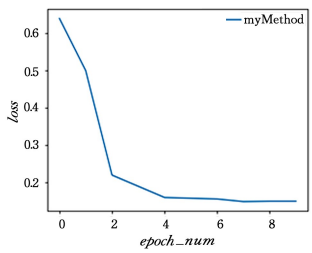
对齐本体	本文方法	LR[10]	RF[10]	XGBoost[10]	MP[9]	REPTree[9]	M5[9]
平均	0.71	0.62	0.64	0.65	0.67	0.65	0.65

本文使用自注意力模型能得到每个相似性特征的注意力权重，通过实验可知，将文中使用的基于字符串、基于语义和基于结构信息这 3 类相似性度量作为特征，其中基于语义的特征的权重相比其他特征的权重更大，说明基于语义的相似性特征对分类结果的贡献更大，这是因为匹配单元对的文本内容在语义上具有更深层次的对应关系。

图 4 给出了本文方法分别在两个数据集上训练的 loss 曲线，其中横坐标为训练轮数（用 *epoc_num* 表示），纵坐标为损失值（用 *loss* 表示），可以看出，随着训练轮数的增加，loss 曲线收敛。



(a) dataset 1



(b) dataset 2

图 4 本文方法分别在两个数据集上的 loss 曲线

Fig. 4 Loss curves of the proposed method on two datasets

图 5 给出了本文方法分别在两个数据集上使用不同 dropout 比率对 F-Measure 的影响，其中横坐标为 dropout 比率（用 dropou trate 表示），纵坐标为 F-Measure 值（用 F-Measure 表示），可以看出，dropout 的使用可以降低过拟合现象，当 $dropout=0.5$ 时，F-Measure 值达到最高，随着 dropout 比率继续增大，训练变得不稳定。

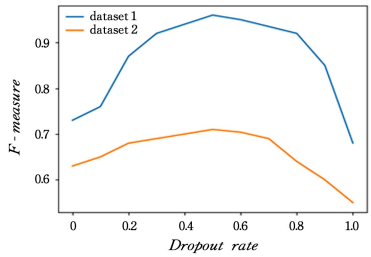


图 5 在两个数据集上 dropout 比率对实验结果的影响

Fig. 5 Effects of dropout ratio on experimental results on two data sets

5.3.4 消融实验

为了进一步验证本文提出的基于 Focal loss 损失函数和 SMOTEENN 方法的有效性，本文进行了消融实验。1）将 5.3.3节中本文方法中的 Focal loss 损失函数换回普通的交叉熵损失函数，记为“本文方法-Focal loss”；2）将 5.3.3 节中本文方法中的 SMOTEENN 步骤去掉，记为“本文方法-SMOTEENN”；3）在“本文方法-SMOTEENN”的基础上，采用原始的交叉熵损失函数，记为“本文方法-SMOTEENN-Focal loss”。表 6 列出了这 3 种方法分别在数据集 1 和数据集 2 上的实验结果。

表 6 数据集 1、数据集 2 上使用不同本体对齐方法的 F-Measure

Table 6 F-Measure in Dataset 1 and Dataset 2 with different ontology alignment methods

对齐本体	本文方法	本文方法-Focal loss	本文方法-SMOTEENN	本文方法-SMOTEENN-Focal loss
数据集 1	0.96	0.82	0.91	0.79
数据集 2	0.71	0.67	0.69	0.63

如表 6 所列，类别不平衡问题给分类结果带来了影响，使用 SMOTEENN 过采样和 Focal loss 损失函数均能改善这个问题。对于两个数据集，单独使用 Focal loss 比单独使用 SMOTEENN 的效果更明显，这是因为改进损失函数能有效地使得负样本的损失变小以及正样本的损失变大，能够更加完整地保留原始数据包含的信息。

结束语 本文提出了一种基于自注意力模型的本体对齐方法，将两个本体中的类和属性分别构建匹配单元对，对每个匹配单元对进行基于字符串、基于语义和基于结构的相似性度量，构建相似性特征向量，然后综合两个本体间的所有匹配单元的相似性特征向量组成相似性特征矩阵，基于该矩阵训练可用于匹配单元对齐的自注意力模型，从而实现本体对齐。通过实验验证了本文方法的有效性，相比其他方法，本文方法的 F-Measure 均有明显提升，这是因为使用自注意力模型能够有效聚合多种不同的相似性度量，以获得更好的对齐结果。在未来的研究中，可以考虑对本体中的类和属性赋予不同的权重，分别进行度量和对齐，进一步提高本体对齐的精度。

参 考 文 献

[1] EUZENAT J, SHVAIKO P. Ontology Matching[M]. Berlin: Springer-Verlag Berlin Heidelberg, 2007: 25-54.

[2] WANG S, KANG D Z, JIANG D Y. Survey of Ontology Mapping[J]. Computer Science, 2017, 44(9): 1-10.

[3] XU J, FANG A, HONG N. An Ontology Mapping Method Based on Lexical Similarity Calculation[J]. New Technology of Library and Information Service, 2013, 29(2): 36-42.

[4] YAO X M, WANG F, LIN L F, et al. An Efficient Multi-policy Ontology Mapping Method[J]. Chinese Science and Technology Papers, 2013, 8(7): 642-647.

[5] YU J, XIONG Z H, OU Z H. Eliminating Redundant Ontology Relations Based on Hasse Diagram[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(3): 279-285.

[6] JIANG M, YU M G, WANG Z X. Multi-strategy Adaptive Large-scale Ontology Mapping Algorithm[J]. Computer Engineering, 2019, 45(3): 14-19.

[7] EUZENAT J, GUÉGAN P, VALTCHEVP. OLA in the OAEI 2005 alignment contest[C]// Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies. 2005: 61-71.

[8] NEZHADI A, SHADGAR B, OSAREH A. Ontology Alignment Using Machine LearningTechniques[J]. International Journal of Computer Science & Information Technology, 2011, 12(3): 139-150.

[9] ALBOUKAEY N, JOUKHADAR A. Ontology Matching as Regression Problem[J]. Journal of Digital Information Management, 2018, 16(1): 85-99.

[10] LEV B, SERGEY S. Applying of Machine Learning Techniques to Combine String-based, Language-based and Structure-based Similarity Measures for Ontology Matching[C]// Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains. 2019: 129-147.

[11] SENGODAN M, SAMUKUTTY A. Explicit Link Discovery Scheme Optimized with Ontology Mapping using Improved Machine Learning Approach[J]. Studies in Informatics and Control, 2021, 30(1): 189-201.

[12] SABOU M, THIÉBLIN E, HAEMMERLÉ O, et al. Survey on complex ontology matching[J]. Semantic Web, 2020, 11(4): 32-62.

[13] WANG R J. Research on Ontology Mapping Methods[D]. Changchun: Jilin University, 2012.

[14] LOU W, WANG H, JU Y. An ontology fusion method based on binary similarity calculation[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(6): 622-631.

[15] SUN X. Research on Ontology Alignment Based on Word Embedding[D]. Harbin: Harbin Institute of Technology, 2020.

[16] CHEN E J, JIANG E B. Review of Studies on Text Similarity Measures[J]. Data Analysis and Knowledge Discovery, 2017, 1(6): 1-11.

[17] LIN H L, WANG Y Z, JIA Y T, et al. Network Big Data Oriented Knowledge Fusion Methods: A Survey[J]. Chinese Journal of Computers, 2017, 40(1): 1-27.

[18] KANG S Z, JI L X, ZHANG J P. Ontology Alignment Method Based on Word Embedding and Conceptual Context Information [J]. Journal of Information Engineering University, 2020, 21(5): 607-613.

[19] SAEDI C, BRANCOA, RODRIGUES J, et al. Wordnet embeddings[C]// Proceedings of the Third Workshop on Representation Learning for NLP. 2018: 122-131.

[20] JANG B, KIM I, KIM J W. Word2vec convolutional neural networks for classification of news articles and tweets[J]. PloS One, 2019, 14(8): 178-189.

[21] LE Q V, MIKOLOV T. Distributed Representations of Sentences and Documents[C]// International Conference on Machine Learning. 2014: 1188-1196.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[23] LIN T, GOYAL P, GIRSHICK R, et al. Dollar Piotr. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(2): 178-185.

[24] DAVID J, GUILLET F, BRIAND H. Association Rule Ontology Matching Approach[J]. In: International Journal on Semantic Web and information systems, 2007, 3(2): 27-49.

[25] ECKERT K, MEILICKE C, STUCKENSCHMIDT H. Improving ontology matching using meta-level learning[C]// European Semantic Web Conference. Berlin: Springer, 2009: 158-172.

WU Zi-yi, born in 1998, master. Her main research interests include knowledge graph and NLP.

LI Shao-mei, born in 1982, Ph.D, associate professor, master supervisor. Her main research interests include knowledge graph and NLP.

(责任编辑:喻藜)