第47卷 第5期 计 算 机 工 程 2021年5月
Vol.47 No.5 Computer Engineering May 2021

·人工智能与模式识别· 文章编号:1000-3428(2021)05-0097-07 文献标志码:A 中图分类号:TP18

基于机器阅读理解模型与众包验证的属性值抽取方法

冯 秒,刘井平,蒋海云,肖仰华

(复旦大学 计算机科学技术学院,上海 200433)

摘 要: 由于互联网语料的高噪音特性,传统的属性值抽取方法存在人工成本增加及训练集缺乏等问题。提出一种新的实体属性值抽取方法。利用机器阅读理解模型,从互联网语料中抽取出高质量的候选属性值,通过高效的众包验证机制调整各候选属性值的权重,得到最终抽取结果。实验结果表明,与OpenTag、QANET等模型相比,该机器阅读理解模型有效提升了候选属性值抽取的准确性,抽取准确率提升10%左右,同时通过众包验证方法,能够以较低的众包成本提高属性值抽取的整体性能。

关键词: 属性值抽取;机器阅读理解模型;知识图谱;众包;序列标注

开放科学(资源服务)标志码(OSID):

中文引用格式: 冯秒,刘井平,蒋海云,等.基于机器阅读理解模型与众包验证的属性值抽取方法[J].计算机工程,2021,47(5):97-103.

英文引用格式: FENG Suo, LIU Jingping, JIANG Haiyun, et al. Attribute value extraction method based on machine reading comprehension model and crowdsourcing verification[J]. Computer Engineering, 2021, 47(5): 97-103.

Attribute Value Extraction Method Based on Machine Reading Comprehension Model and Crowdsourcing Verification

FENG Suo, LIU Jingping, JIANG Haiyun, XIAO Yanghua

(School of Computer Science, Fudan University, Shanghai 200433, China)

【Abstract】 Due to the high noise characteristics of Internet corpus, traditional extraction methods based on attribute values suffer from increased labor costs and lack of training sets. This paper proposes an entity attribute value extraction method based on machine reading comprehension model and crowdsourcing verification. The new machine reading comprehension model is used to extract high-quality candidate attribute values from the Internet corpus, and the weight of each candidate attribute value is adjusted through an efficient crowdsourcing verification mechanism to obtain the final extraction result. Experimental results show that compared with OpenTag, QANET and other models, the machine reading comprehension model effectively improves the accuracy of candidate attribute value extraction, and the extraction accuracy is increased by about 10%. At the same time, it can improve the overall performance of attribute value extraction at a low crowdsourcing cost by using crowdsourcing verification.

【Key words】 attribute value extraction; machine reading comprehension model; knowledge graph; crowdsourcing; sequence labeling

DOI: 10.19678/j.issn.1000-3428.0057666

0 概述

近年来,随着自然语言处理需求的不断增长,知识图谱成为人们研究的焦点。目前,知识图谱已经被广泛地应用于各种智能化的推荐系统[1]、问答系统[2]以及语言生成[3]等诸多场景。为满足这些应用的数据需求,研究人员构建出大规模的知识图谱,如英文的DBpedia[4]、YAGO[5]以及中文的CN-DBpedia[6]等。在这些知识图谱中,数据通常以<实体,属性(关系),属性值(尾实体)>的三元组形式进行组织,例如“特朗普的国籍是美国”这一条知识,就可以被表示为<特朗普,国籍,美国>。由于本文所叙述的方法对属性值及尾实体的抽取均适用,后文将属性和关系统一称为属性,将属性值和尾实体统一称为属性值。

基金项目: 上海市科技创新行动计划(19511120400)。

作者简介: 冯 秒(1994—),男,硕士研究生,主研方向为知识图谱、信息抽取;刘井平、蒋海云,博士研究生;肖仰华,教授、博士。

收稿日期: 2020-03-10 **修回日期:** 2020-04-17 **E-mail:** fengs17@fudan.edu.cn

计算机工程 www.ecice06.com

在构建知识图谱的过程中，获取图谱中实体相关的各属性的属性值是非常重要的步骤。传统的属性值抽取方法主要分为基于模式的方法、基于分类器的方法和基于序列标注模型的方法。基于模式的方法通常被用于从高质量的半结构化文本（如百科页面的信息表）中，利用句法模式[7]、语义模式[8]或正则表达式等人工或自动生成的模式直接进行抽取。由于这类半结构化文本的表达规律性强，通过基于模式的方法可以用较小代价得到大量高质量的三元组。此类方法已经被应用于大量的知识图谱构建实践中。基于分类器的方法首先需要利用命名实体识别（NER）工具，在文本中定位整个三元组的实体和属性值。然后利用类似PCNN[9]及文献[10]所述的分类器，结合实体和属性值之间的上下文信息，分类得到三元组对应的属性，从而实现抽取。基于序列标注模型的方法最初的序列标注模型常被应用于NER等任务中[11]。随着序列标注模型的发展，序列标注模型也开始被应用于抽取任务，例如，人们利用BiLSTM-CRF神经网络实现了已知属性下的实体和属性值抽取[12]。近年来，随着注意力机制的不断发展，人们也开始使用更强的序列标注模型，实现了具有较高准确率的属性值抽取方法[13]。

随着在真实场景下数据需求的不断扩展，知识图谱开始需要更多细节的知识，这使得从互联网语料中抽取更多更稀有的属性成为知识图谱构建中的新问题。传统的基于模式的方法由于互联网语料的高噪音，导致人工构建模式的工作量大幅增加，自动构建模式准确率下降，难以召回足够的高质量知识。而对于和训练集有着较强相关性的分类方法和序列标注方法，目前仍然缺乏高质量的基于互联网语料的训练数据。并且由于模型本身的限制，它们难以抽取超出训练集之外（缺乏训练数据）的稀有属性，使得其应用受到了限制。

考虑到传统方法在新场景下所受到的限制，本文提出一种基于机器阅读理解模型和众包验证的属性值抽取方法。该方法的输入为所需抽取的实体-属性对，通过构造搜索关键字从互联网中获得上下文，并根据机器阅读理解模型，以类似序列标注模型的形式标注出候选属性值并得到置信度。在此基础上，通过轻量化的众包验证方法，对模型得到的抽取结果进行验证，并调整优化各候选结果置信度以得到最终的抽取结果。

1 本文方法架构

实体属性值抽取任务的输入为所需抽取的实体-属性对，输出为抽取得到的属性值。本文属性值抽取方法总体框架如图1所示。

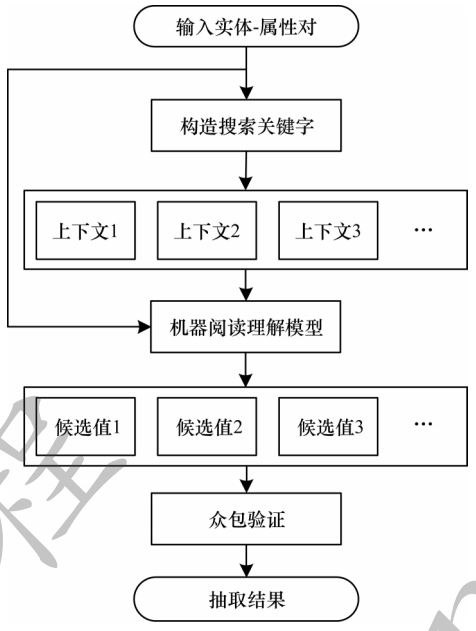

图1 本文方法的总体框架

Fig.1 Overall framework of the proposed method

本文方法主要步骤如下：

步骤1 构造搜索关键字与获取互联网上下文。对于输入的实体-属性对，需要其相关上下文作为抽取时的知识来源。通过简单的模板，“[实体]的[属性]”可以构造出用于搜索引擎搜索的关键字，例如对于实体-属性对<西虹市首富，上映时间>，可以构造出搜索关键字“西虹市首富的上映时间”。利用上面构造的关键字进行搜索，可以从搜索引擎的结果页中抓取各条结果的摘要信息，由于这些摘要信息为基于实体-属性对构造出的关键词的相关内容，本文直接把搜索结果中的前k条作为获取的互联网上下文。

步骤2 基于机器阅读理解模型的候选属性值抽取。机器阅读理解模型以实体-属性对和一条对应的上下文作为输入，并输出从该上下文中抽取出的候选属性值及其置信度。

步骤3 众包验证。得到候选属性值及其置信度后，本文引入一种基于判断题形式的众包验证任务。在人工标注中，仅需要根据提示对各候选属性值进行简单判断。通过人工判断的结果对上面步骤得到的候选属性值的置信度进行调整，最终选择置信度最高的一条结果作为输出。

2 机器阅读理解模型

为实现对候选属性值的抽取，本文提出了针对属性值抽取的机器阅读理解模型，即MCKE（Machine reading Comprehension Knowledge Extraction）模型。该模型主要来自文献[14-15]中的两个基于注意力机制的机器阅读理解模型，它们在机器阅读理解任务上有着较高的准确率。针对属性值抽取任务，MCKE模

计算机工程 www.ecice.com

型通过引入BERT进行字级别的输入表示学习，并在输入表示的过程中进行了上下文表示的增强，从而提高其对属性值抽取任务的适应性。MCKE模型的整体架构如图2所示。

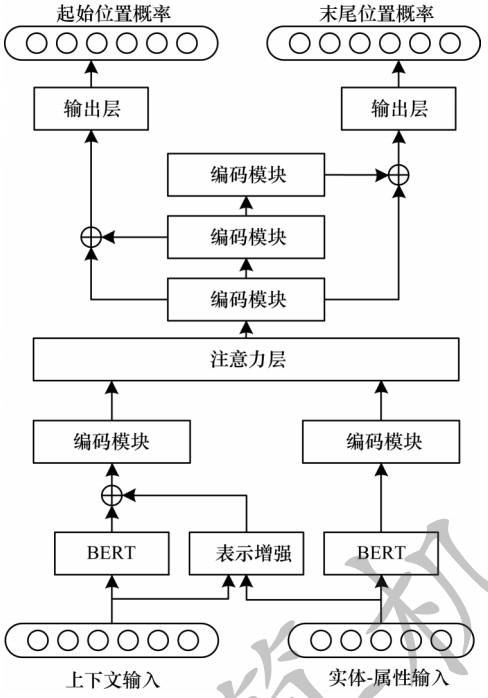

图2　MCKE模型的整体架构

Fig.2　Overall framework of MCKE model

本文模型的输入为上下文字符串C，以及用空格隔开的由实体和属性拼接而成的字符串Q。模型的输出为最终抽取结果的起始位置以及末尾位置的概率分布，分别为p^s和p^e。

本文模型主要包含输入表示、表示变换、注意力层、输出层等组成部分。

2.1　输入表示

2.1.1　字符级别表示

不同于常用的Word2Vec[16]、GloVe[17]等词嵌入方法，本文选择使用Google最新提出的BERT[18]语言模型来获得模型输入部分的向量表示。BERT模型是一种Transformer[19]结构实现的语言模型，实现了对于各输入字符上下文的编码。不同于传统词嵌入方法中字符的表示与上下文无关，本文引入BERT可以使同一个字符输入在上下文不同的情况下有着不同的向量表示，其包含了与字符同时输入的上下文信息。

在MCKE模型中使用了预训练的BERT模型来获得输入信息的向量表示，对于实体-属性输入和上下文输入$Q=\{q_1,q_2,\cdots,q_m\}$和$C=\{c_1,c_2,\cdots,c_n\}$，先在其首尾添加标识符[CLS]和[SEP]，将其输入BERT模型中，得到输入实体-属性表示与上下文表示的$U_0=\{u_1,u_2,\cdots,u_m\}$和$H_0=\{h_1,h_2,\cdots,h_n\}$。

2.1.2　增强表示

除字符级别的表示外，本文针对上下文输入进行了表示的增强。通过计算额外信息向量，在表示中标注了和实体-属性有关部分的信息，也间接补充了分词的信息。在上下文输入中引入实体-属性输入内容的位置信息，可有助于模型对句子结构信息的学习，增强其对稀有属性的抽取效果，增强表示方法如下：

1）对输入的上下文和实体-属性进行分词得到$C_{\text{word}}=\{cw_1,cw_2,\cdots\}$和$Q_{\text{word}}=\{qw_1,qw_2,\cdots\}$，其中，$qw_i$和$cw_i$为分词结果得到的词语。利用这部分信息，可以计算上下文中各词的额外信息向量$\boldsymbol{cw}_i^+$，其组成部分如式（1）所示：

$$\boldsymbol{cw}_i^+=\left[|\{\boldsymbol{cw}_i^+\}\cap Q_{\text{word}}|,\frac{|\boldsymbol{cw}_i^+\cap Q|}{|\boldsymbol{cw}_i^+|}\right] \tag{1}$$

2）每个信息向量包含两个维度，其中，第一维表示该词语是否被包含在实体-属性输入中，第二维表示该词语中各字符被包含在实体-属性输入中的比率。计算得到各额外信息向量后，将其与H_0中各对应的字符向量进行拼接，得到增强过的上下文表示$\overline{H_0}$，如式（2）所示：

$$\overline{H_0}=\{[h_1,cw_1^+],[h_2,cw_2^+],[h_3,cw_3^+],\cdots\} \tag{2}$$

2.2　变换表示

基于本文提出的模型，在整个模型的输入和输出过程中需要对文本的表示进行变换，从而生成其后面步骤所需的表示。为此，本文引入了类似于文献[14]所采用的编码模块，并通过这些编码模块实现表示的变换。

在一个编码模块中，对于一个输入表示X，首先对该表示进行位置编码。位置编码是一种常用的用于在非循环神经网络中保留输入位置信息的方法。经过位置编码后，得到新的表示$\hat{X}$。

在编码模块中，使用了3种不同的编码层对输入表示进行变换，即卷积层、自注意力层和前馈层。

编码模块中的卷积层选用了Separable卷积层[20]，其相比普通的卷积层有着计算速度快、泛化能力强的优势。在本文模型中，卷积核大小设为7，并通过padding方法补全输出至和输入相同维度。

编码模块中的自注意力层使用了多头注意力的方式来计算，从而增强自注意力层对于多关注点的信息，对于输入X进行如下计算：

$$X_i=XW_i^X \tag{3}$$

$$\text{head}_i=\text{Attention}(X_i,X_i,X_i) \tag{4}$$

$$O=\text{Concat}(\text{head}_1,\text{head}_2,\cdots)W^O \tag{5}$$

其中，O为本层输出，W_i^X和W^O为可训练的参数。本文模型自注意力层的头数设为2。

编码模块中的前馈层计算方法如下：

$$O=\text{relu}(W^FX)W^{FO} \tag{6}$$

计算机工程 www.ecjo06.com

为防止训练过程中可能出现的梯度消失现象，在上述各层的使用中都加入了残差机制。对于编码模块中各层，其输出为：

$$O=\text{Layer}(\text{layernorm}(X))+X \tag{7}$$

其中，layernorm为Layer Normalization[21]。

通过上述3种层的结合，可以得到2种不同的编码模块，在注意力层之前的编码模块，由4个卷积层、1个自注意力层和1个前馈层组成。在注意力层之后的编码模块，由2个卷积层、1个自注意力层和1个前馈层组成。

2.3 注意力层

注意力层的输入是前面步骤获得的上下文表示和实体-属性表示，经过编码模块之后得到的新表示结果 $H_1=\text{Encoder}(\overline{H}_0)$ 和 $U_1=\text{Encoder}(U_0)$。在本文步骤中，主要利用注意力机制对上下文表示进行进一步的增强。

采用类似于文献[15]的方法，本文模型需要首先通过式(8)中的线性变换，得到实体-属性表示的相似性矩阵 $\boldsymbol{S}$。

$$S_{ij}=f(u_i,h_j)=W^s[u_i,h_j,u_i\odot u_j] \tag{8}$$

其中，$\odot$ 为元素级别的乘法，W^s 为可训练的参数。利用相似性矩阵 $\boldsymbol{S}$，经过式(9)~式(12)中的变换，可以进一步计算上下文对实体-属性对的注意力 $\boldsymbol{A}$，以及实体-属性对与上下文的注意力 $\boldsymbol{B}$。

$$\hat{\boldsymbol{S}}=\text{softmax}(\boldsymbol{S},\text{axis}=\text{row}) \tag{9}$$

$$\bar{\boldsymbol{S}}=\text{softmax}(\boldsymbol{S},\text{axis}=\text{column}) \tag{10}$$

$$\boldsymbol{A}=\hat{\boldsymbol{S}}\cdot\boldsymbol{U}^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{B}=\hat{\boldsymbol{S}}\cdot\bar{\boldsymbol{S}}^{\mathrm{T}}\cdot\boldsymbol{H}^{\mathrm{T}} \tag{12}$$

其中，$\hat{\boldsymbol{S}}$ 和 $\bar{\boldsymbol{S}}$ 分别为利用softmax对相似性矩阵 $\boldsymbol{S}$ 按行和按列进行归一化的结果。

最终通过对上述注意力以及表示的组合，得到融合了实体-属性信息的新上下文表示 G，其中 $G_i=[h_i,a_i,h_i\odot a_i,h_i\odot b_i]$。

2.4 模型输出

经过注意力层得到的新的上下文表示 G 后，经过3个额外的编码模块，可以分别得到用于计算输出的3个新的上下文表示：$G_0=\text{Encoder}(G)$，$G_1=\text{Encoder}(G_0)$，$G_2=\text{Encoder}(G_1)$。

模型的输出层通过对拼接之后的编码进行一次非线性变化，可以得到抽取结果的起始位置和末尾位置在上下文输入中位置的概率分布：

$$p^s=\text{softmax}(W^1[G_0;G_1]) \tag{13}$$

$$p^e=\text{softmax}(W^2[G_0;G_2]) \tag{14}$$

其中，W^1 与 W^2 为可训练的参数，[;]为向量拼接操作，p^s 为起始位置分布，p^e 为末尾位置分布。

本文模型的目标函数为：

$$L(\theta)=-\frac{1}{N}\sum_{i}^{N}[\log_a(p^s_{y_i^s})+\log_a(p^e_{y_i^e})] \tag{15}$$

其中，y_i^s 与 y_i^e 分别对应第 i 个样本中真实结果起止位置所在的坐标。在训练过程中，通过最小化该目标函数使得真实结果坐标处的概率最大化。

2.5 抽取结果的获取

基于模型输出的起始位置分布 p^s 以及末尾位置分布 p^e，通过最大化 $p^s_{\text{start}}p^e_{\text{end}}$ 得到相应的起始和终止位置(start, end)，就可以得到最终的抽取结果所在的坐标。该抽取结果的置信度为 $\omega_i=p^s_{\text{start}}p^e_{\text{end}}$。

3 众包验证

在经过模型抽取后可以发现，一些容易混淆的结果通常在置信度上相差不多。本文参照文献[22]中的架构，引入一种高效率的判断题式众包验证任务，对合理的抽取结果进行奖励，对不合理的结果进行惩罚来进一步提高抽取的效果。

3.1 众包任务设计

本文将众包任务设定为判断题，以提高众包任务的时间效率。将抽取任务中的上下文、实体-关系对以及抽取的候选属性值展示给众包工人，由工人判断这一候选属性值是否合理。通过这样的手段，降低了工人的准入门槛，提高了任务的完成效率。本文所设计的众包任务的实际界面如图3所示。

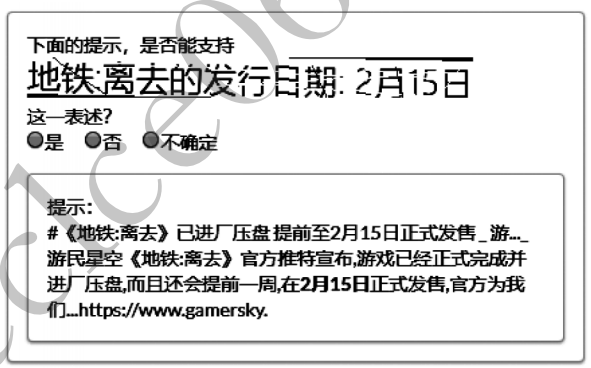

图3 众包验证的用户界面

Fig.3 User interface for crowdsourcing validation

在界面中，为方便定位，在上下文中对候选属性值进行了加粗处理，方便在遇到困难情况时快速跳转到下一条任务，提供不确定的选项。

3.2 结果权重调整

通过对各结果进行众包后，可以获得对每条候选结果的包含正确、错误及不确定3种选项的信任度投票情况。基于这些情况，对在众包验证中被选择正确占比较高的候选结果的权重进行奖励，对被选择错误数量占比较高的结果的权重进行惩罚，并保持被选择不确定的结果的权重。可以区分置信度相差不大的易混淆结果，并对模型抽取出的错误结果进行排除。参考文献[23]中的基于正确、错误、不确定3项标注的置信度计算方法，可利用式(16)计算得到调整系数 γ_i：

www.ecice06.com

$$\gamma_i = \frac{2n_i^{\text{pos}} + n_i^{\text{neu}} + 1}{n_i^{\text{pos}} + 2n_i^{\text{neg}} + n_i^{\text{neu}} + 1} \tag{16}$$

$$\omega_i' = \gamma_i \cdot \omega_i \tag{17}$$

其中,n_i^{pos}、n_i^{neg}和n_i^{neu}分别表示对第i条候选属性值选择正确、错误、不确定的众包验证结果数量。利用该调整系数对之前计算所得的结果权重根据式(17)进行调整,可以获取各候选属性值的调整后分数ω_i'。考虑到在进行抽取中,不同上下文可能抽取得到相同的候选属性值,在最终计算过程中,对于这些相同属性值的调整分数进行求和,得到最终的抽取分数。

最终以具有最高抽取分数的属性值作为抽取结果输出,其中分数大于阈值β的结果被作为有效抽取结果。

4 实验

4.1 实验数据

本文所采用的实验数据分为2个部分:1)抽样数据,从CN-DBpedia中抽样得到的162组实体-属性对以及其对应的属性值;2)新实体数据,采用人工标注得到的496组实体-属性对以及对应属性值。为保证实验数据的多样性,在上面的实验数据获取过程中,选择了分别来自于手机、电影、事件、游戏、显卡等多种类别实体。

在实验的过程中,上下文信息使用了在搜索引擎(百度)上的结果页面,将召回的各条结果信息分条作为实体-属性对的上下文。对于每组搜索关键字获取的结果数量k设为10。

模型的训练数据包括以下2个部分:1)来自文献[24]所提供的中文机器阅读理解数据集;2)通过远程监督方法从CN-DBpedia和百度搜索引擎中得到的数据集。

4.2 候选属性值抽取分析

4.2.1 实验设置

为验证本文所提出的候选属性值抽取模块(机器阅读理解模型)的有效性,选择了另外3种模型作为基线。

1)序列标注模型OpenTag[13]

OpenTag是最新的一种适用于开放域属性值抽取的基于自注意力机制的序列标注模型,其在多个数据集上有着超过常用的BiLSTM以及BiLSTM-CRF模型的效果。

2)机器阅读理解模型QANET[14]

QANET是由Google Brain所提出的基于注意力机制的机器阅读理解模型,它长期在机器阅读理解的SQuAD数据集的排行榜上位置靠前。

3)去除表示增强的MCKE-NA模型

为验证表示增强部分的效果,本文也引入了不包含表示增强部分,仅使用字符级表示作为输入的MCKE模型进行对比。

为评测模型本身的效果,在最终抽取结果的选定上,使用了一种简单策略来计算最终的候选属性值分数,类似于2.2节所述,对有多个来源的候选属性值,令其抽取分数为与之相同的所有候选属性值的置信度之和。最后,选取分数最高的候选属性值作为输出,对于模型的输出结果,其阈值设为$\beta = 0.1$。

4.2.2 实验结果与分析

表1列出了各基线模型与本文MCKE模型在两组数据上通过执行上述简单策略进行属性值抽取的结果,并分别评测了抽取结果的准确率以及抽取结果PR曲线(Precision-Recall曲线)的曲线下面积(AUC)。

表1 属性值抽取模型效果

Table 1 Effect of attribute value extraction models

模型	抽样数据		新实体数据	
	准确率	AUC(PR曲线)	准确率	AUC(PR曲线)
OpenTag	0.280	0.039	0.240	0.043
QANET	0.327	0.220	0.438	0.126
MCKE-NA	0.333	0.184	0.353	0.119
MCKE	0.457	0.340	0.599	0.275

从表1可以看出,本文的MCKE模型在两组数据中均在准确率和AUC上超过了所选的对比基线。在实验中发现,基于标注模型的OpenTag方法的表现受限于训练数据的结构,难以召回训练数据中较为稀有的属性所对应的属性值,导致其召回率的下降。并且,由于OpenTag仅返回标注结果,难以使用阈值对结果进行筛选,导致了其在准确率上表现偏低,对于QANET和MCKE-NA两种机器阅读理解模型,受限于它们本身对表示增强的缺乏,使得它们在稀有属性中效果偏低,而MCKE-NA由于缺乏词级别的表示学习,导致其结果属性值不完整的情况较多。

表2列出了实验中2种机器阅读理解模型得到的抽取结果。从表2可以看出,对于显卡相关的在知识库中较为稀有的属性,通过使用机器阅读理解模型可以实现抽取,并且MCKE模型的性能更强。同时,对于电影相关属性,得益于MCKE模型所采用的更优秀的输入表示学习方法与表示增强方法,可以抽取出语义更加完整的属性值结果。

计算机工程 www.ecice06.com

表2 属性值抽取模型结果实例

Table 2 Result examples of attribute value extraction models

实体	属性	QANET结果	MCKE结果
飞驰人生	拍摄地点	新疆	新疆巴音布鲁克
飞驰人生	对白语言	汉语	汉语普通话
TITAN V	显存位宽	3 072 bit	3 072 bit
TITAN V	TDP	250 W	250 W
TITAN V	架构	NVIDIA	Volta

4.3 众包验证分析

4.3.1 实验设置

为验证众包验证步骤的有效性,在本文实验中,将每个通过MCKE模型得到的候选属性值及其相关上下文输入到众包验证模块,并邀请了5位志愿者,每人随机完成150个众包任务。对众包结果进行统计后调整权重,得到最终的抽取结果。

4.3.2 众包验证分析结果

表3列出了随机进行750个任务以及其中前200个任务,在众包验证之后所实现的抽取效果。通过对总共750个众包任务的日志进行分析,发现本文众包任务的平均响应时间为4.21 s。

表3 众包验证效果

Table 3 Effect of crowdsourcing validation

任务	准确率	AUC(PR曲线)
MCKE	0.599	0.275
MCKE+200条众包验证	0.641	0.308
MCKE+750条众包验证	0.688	0.345

实验结果表明,在增加了众包验证的步骤后,在其他实验设置相同的情况下,可以纠正MCKE模型抽取结果中的错误,使得准确率大幅上升,并且可以发现,其中存在的一些由于模型缺乏训练数据所导致的常识性错误结果以及其中出现了边界错误的结果,可以被较快纠正,如表4所示。

表4 众包验证结果实例

Table 4 Result examples of crowdsourcing validation

实体	属性	原结果	众包验证后结果
华为P11	处理器	6 GHz	麒麟970处理器
华为P11	ROM	麒麟970处理器	N/A
iPhone XS	芯片	A1	A12

另外,在进行分析的过程中发现,通过调整置信度阈值,本文方法在阈值β为0.4的情况下抽取出属性值的准确率为0.857。

5 结束语

本文针对日益增长的知识图谱数据需求,提出一种基于机器阅读理解以及众包验证的实体属性值抽取方法。从互联网获取相关上下文,利用机器阅读理解模型从相关上下文中获取候选属性值,通过众包验证从候选属性值中找出最优的抽取结果。实验结果表明,该框架能够有效地从互联网语料中抽取出实体属性值,且提出的众包任务能够以较高的效率提升整体抽取效果。本文对于众包结果的应用仅针对抽取模型的结果验证,而从众包结果中获取的其他信息,也可成为模型增强训练的重要反馈,对于该部分信息的应用将是下一步的研究重点。

参考文献

[1] WU Xiyu, CHEN Qimai, LIU Hai, et al. Collaborative filtering recommendation algorithm based on representation learning of knowledge graph[J]. Computer Engineering, 2018, 44(2): 226-232, 263. (in Chinese)
吴玺煜,陈启买,刘海,等. 基于知识图谱表示学习的协同过滤推荐算法[J]. 计算机工程,2018,44(2):226-232,263.

[2] DENG Yang, XIE Yuexiang, LI Yaliang, et al. Multi-task learning with multi-view attention for answer selection and knowledge base question answering[EB/OL]. [2020-02-08]. https://arxiv.org/abs/1812.02354.

[3] CHEN Jiangjie, WANG Ao, JIANG Haiyun, et al. Ensuring readability and data-fidelity using head-modifier templates in deep type description generation[EB/OL]. [2020-02-08]. https://arxiv.org/abs/1905.12198.

[4] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: a nucleus for a Web of open data[C]//Proceedings of the 6th International Semantic Web Conference. Berlin, Germany: Springer, 2007: 722-735.

[5] REBELE T, SUCHANEK F, HOFFART J, et al. YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames[C]//Proceedings of the 15th International Semantic Web Conference. Berlin, Germany: Springer, 2016: 177-185.

[6] XU Bo, XU Yong, LIANG Jiaqing, et al. CN-DBpedia: a never-ending Chinese knowledge extraction system[C]// Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Berlin, Germany: Springer, 2017: 428-438.

[7] HWARST M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th Conference on Computational Linguistics. Stroudsburg, USA: ACL, 1992: 539-545.

[8] CHEN Jindong, WANG Ao, CHEN Jiangjie, et al. CN-Probase: a data-driven approach for large-scale Chinese taxonomy construction[C]//Proceedings of the 35th International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2019: 1706-1709.

[9] JIANG Xiaotian, WANG Quan, LI Peng, et al. Relation extraction with multi-instance multi-label convolutional neural networks[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Stroudsburg, USA: ACL, 2016: 1471-1480.

[10] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings

计算机工程 www.ecice.com

of Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, USA: ACL, 2009: 1003-1011.

[11] ZHANG Yingcheng, YANG Yang, JIANG Rui, et al. Commercial intelligence entity recognition model based on BiLSTM-CRF[J]. Computer Engineering, 2019, 45(5): 308-314. (in Chinese)
张应成,杨洋,蒋瑞,等. 基于 BiLSTM-CRF 的商情实体识别模型[J]. 计算机工程,2019,45(5):308-314.

[12] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2016: 1105-1116.

[13] ZHENG G, MUKHERJEE S, DONG X L, et al. Opentag: open attribute value extraction from product profiles[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2018: 1049-1058.

[14] YU A W, DOHAN D, LUONG M T, et al. Qanet: combining local convolution with global self-attention for reading comprehension[EB/OL]. [2020-02-08]. https://openreview. net/forum? id=B14TlG-RW.

[15] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. [2020-02-08]. https://openreview. net/forum? id=HJ0UKP9ge.

[16] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of NIPS'13. New York, USA: ACM Press, 2013: 3111-3119.

[17] PENNINGTON J, SOCHER R, MANNING C D. GloVe: Global vectors for word representation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2014: 1532-1543.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the 9th American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: ACL, 2019: 4171-4186.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of NIPS'13. New York, USA: ACM Press, 2017: 5998-6008.

[20] KAISER L, GOMEZ A N, CHOLLET F. Depthwise separable convolutions for neural machine translation [EB/OL]. [2020-02-08]. https://openreview. net/forum? id=S1jBcueAb.

[21] BA J L, KIROS J R, HINTON G E. Layer normalization [EB/OL]. [2020-02-28]. https://arxiv. org/abs/1607. 06450.

[22] KONDREDDI S K, TRIANTAFILLOU P, WEIKUM G. HIGGINS: knowledge acquisition meets the crowds[C]// Proceedings of the 22nd International Conference on World Wide Web. New York, USA: ACM Press, 2013: 85-86.

[23] LI Qi, JIANG Meng, ZHANG Xikun, et al. Truepie: discovering reliable patterns in pattern-based information extraction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2018: 1675-1684.

[24] LI Peng, LI Wei, HE Zhengyan, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering [EB/OL]. [2020-02-08]. https://arxiv. org/pdf/1607. 06275. pdf.

编辑　索书志

计算机工程 ce06.com