融合外部知识的生成式实体关系联合抽取方法

祝振赫[1],武 虹[2],高 洁[2],周 玉[3,4]
(1. 中国科学院大学 人工智能学院,北京 100049;
2. 中国科协创新战略研究院,北京 100038;
3. 中国科学院自动化研究所 模式识别国家重点实验室,北京 100190;
4. 北京中科凡语科技有限公司 凡语 AI 研究院,北京 100190)

摘 要:实体关系联合抽取作为各领域构建知识图谱不可或缺的任务,成为当今信息抽取任务中的热点。现有的生成式实体关系联合抽取方法,多采用编码器-解码器框架,通过监督学习从非结构化文本中抽取特征来生成实体和关系序列。但这种方法属于数据驱动方法,在缺乏标注数据时存在质量较低的问题,而获取标注数据需要花费大量的成本。基于远程监督的方法通过利用外部知识库对文本进行自动标注,能够解决缺少大规模标注数据的问题,但同时引入的错误标签也会影响模型的性能。针对上述问题,提出了融合外部知识的生成式实体关系联合抽取方法,采用多编码器和知识注意力机制,将结构化信息和句法结构等外部知识融入模型。具体来说,首先利用标注数据对模型进行预训练来学习实体关系表示,然后利用外部知识再次训练来学习句法结构等信息。实验结果表明,所提方法能够通过融合外部知识,提升实体关系三元组的准确率,尤其提升模型在标注数据稀缺场景下的抽取准确率。

关键词:实体关系抽取;编码-解码框架;知识融合;深度学习;注意力机制
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2023)08-0124-07
doi:10.3969/j.issn.1673-629X.2023.08.018

A Generative Entity Relation Extraction Method Based on External Knowledge

ZHU Zhen-he[1],WU Hong[2],GAO Jie[2],ZHOU Yu[3,4]
(1. School of Artificial Intelligence,University of Chinese Academy of Sciences,Beijing 100049,China;
2. National Academy of Innovation Strategy,Beijing 100038,China;
3. Institute of Automation,Chinese Academy of Sciences,Beijing 100190,China;
4. Fanyu AI Research,Beijing Fanyu Technology Ltd.,Beijing 100190,China)

Abstract:As an indispensable task to construct knowledge map in various fields,entity-relation joint extraction has become a hot topic in information extraction. The existing generative entity-relation joint extraction methods use the encoder-decoder framework to generate sequences of entity and relation. However,this method is data-driven,which has the problem of low quality in the absence of annotation data,and it costs a lot to obtain annotation data. The method based on distant supervision can solve the problem of lack of large-scale annotation data by using external knowledge base to automatically annotate text,but the introduction of wrong labels will also affect the performance of the model. To solve these problems,we propose a generative entity-relation joint extraction method based on external knowledge. The multiple encoders and knowledge attention mechanism is used to fuse the external knowledge such as structured information and syntactic structure to the model. Specifically,the model is first pre-trained with annotated data to learn entity relation representation,and then retrained with external knowledge to learn syntactic structure and other information. The experiment shows the proposed method can improve the accuracy of entity relationship triples by fusing external knowledge,especially the extraction accuracy of the model in the context of label data scarcity.

Key words:entity relation extraction;encoder-decoder frame;knowledge fusion;deep learning;attention mechanism

收稿日期:2022-10-19 **修回日期**:2023-02-22
基金项目:国家自然科学青年科学基金项目(62006224)
作者简介:祝振赫(1995-),男,硕士,研究方向为自然语言处理、知识图谱;通信作者:周 玉(1976-),女,博士,研究员,研究方向为自然语言处理、知识图谱和文本挖掘。

0 引　言

随着互联网、大数据和知识图谱等技术的发展,各行各业都在构建自己的知识库,以便为下游的问答、摘要等应用提供知识的支持。文本作为重要的信息载体,包含了大量丰富的知识,如百科介绍、医院病例、企业标书等,都包含了丰富的内容,具有重要的应用价值。但由于没有统一规范化的写作标准,导致内容形式多种多样且多为非结构化或半结构化文本,为整理统计的工作带来很大的挑战,并难以直接应用于下游任务中。如何从这类非结构化文本中获取蕴含的知识成为了一项重要的研究课题。信息抽取(Information Extraction,IE)是自然语言处理领域中一个非常重要的任务。该任务主要是从非结构化或半结构化文本中抽取出用户所需要的信息,如命名实体、实体关系及事件等,并以结构化的形式输出,以便于支持下游任务的应用[1]。

关系抽取(Relation Extraction,RE)作为信息抽取的主要子任务之一,主要目标是从非结构化文本中抽取出有关的两个实体及实体间的关系,构成关系三元组的结构化形式。通常形式化描述为 $< e_1, r, e_2 >$,其中 e_1、e_2 为实体,r 为两实体间所具有的关系,例如句子"古龙的本名是熊耀华",其中实体"古龙"和实体"熊耀华"具有"本名"的关系[2]。为了区分两个实体,称 e_1 为头实体,e_2 为尾实体。关系抽取是构建知识库的重要基础任务之一,一直受到研究者们的重视。

基于传统的机器学习的关系抽取方法主要通过领域专家制定实体关系范式,通过统计和规则等方法进行抽取。Miller 等人[3]通过训练增强解析树来表示句子的语义信息,再对不同关系类型设计模式匹配规则来进行抽取。Zelenko 等人[4]用核函数及支持向量机的方法进行关系抽取。传统的关系抽取方法具有准确率高的优点,当文本满足规则条件后,就能够抽取出想要的信息。但传统的关系抽取方法依赖人工设计特征和抽取规则,需要大量成本,并难以覆盖所有情况,因此该方法的泛化能力较差。

随着深度学习的兴起,大量的深度学习方法被应用到了关系抽取中。许多经典的关系抽取方法都是使用监督学习来获得较好的性能表现,因为监督学习能够更有效地让模型抽取到特征,从而提高准确率和召回率。但是在一些特定领域,由于难以获取大规模标注数据,导致监督学习的成本大大增加,在应用中通常使用基于启发式规则的无监督方法和远程监督等半监督方法。

网络中的文本形式多样,许多的半结构化文本都包含着一定的实体和关系的知识。如百科类网页大多都详细介绍了一个实体的信息,通常都包含关于该实体的一段描述及其基本信息表。基本信息表一般被称为 Infobox,其中包含了许多和该实体有关的属性和其他实体。这种半结构化文本也经常作为构建知识图谱和远程监督实体关系抽取数据集的直接数据来源[5]。许多基于远程监督的实体关系抽取方法都通过 Infobox 构建的知识图谱来实现。如百度百科"战狼"页面的基本信息中的句子"《战狼》是由吴京执导的现代军事战争片,该片由吴京、余男、倪大红、斯科特·阿金斯、周晓鸥等主演",可以和 Infobox 中的"导演:吴京"条目构成一个实体关系抽取任务的标签数据,原句的一个实体关系元组为<战狼,导演,吴京>。像这种从表格文本以及通过句法规则获取的实体关系三元组都包含了一定的内在知识,如何利用这些内在知识对于实体关系联合抽取十分重要。

该文提出一种基于深度学习的融合外部知识的生成式实体关系联合抽取方法,对于一些难以获取大规模标注数据的特定领域,通过在训练时引入一些典型规则所抽取的结果作为额外知识帮助模型从文本中抽取到更准确全面的信息。模型采用多编码器框架,采用知识自注意力机制对外部知识进行编码。在解码过程中,采用跨知识注意力方法,融合知识编码器的隐层向量得到解码隐层状态,进而生成实体和关系序列。实验结果表明,所提方法能够通过融合外部知识,提升实体关系三元组的准确率,尤其提升模型在标注数据稀缺场景下的抽取准确率。

1 相关工作

关系抽取任务从领域上可分为限定域关系抽取和开放域关系抽取,其区别在于是否对关系类别进行制定。限定域关系抽取需要事先确定关系类别,而开放域关系抽取的关系直接从原文本中抽取获得。限定域关系抽取的优势是所抽取的关系明确,但是无法抽取到所制定关系类别以外的关系。开放域关系抽取能够抽取到更多的关系类别,但经常会抽取到不准确或无意义的关系。

基于深度学习的关系抽取方法主要分为流水线(Pipeline)式抽取方法和联合(Joint)式抽取方法两种。流水线式方法一般是先对文本进行命名实体识别,找出文本中所有的命名实体,然后再将这些命名实体两两配对进行关系分类。这种流水线抽取方法通常用于限定域关系抽取,其关系预测阶段是一个多分类任务,所以无法抽取到定义好的关系以外的其他关系。并且由于分成了两个任务,在实体识别时往往会依赖其他自然语言处理工具,从而导致误差传播的问题,对模型性能会造成一定的影响。当时普遍认为流水线抽取方法不如联合抽取方法好,但 Zhong 等人[6]在流水线模

型的基础上使用了一种简单的方法就达到了当时的最好性能。他们在实体识别结束后对原句进行重构,将抽取到的头尾实体及其类别信息加入到句子当中,然后进行关系分类的判别。可见在关系抽取时引入额外信息有助于模型的学习。联合抽取方法是同时完成实体和实体间关系的抽取任务,通过利用实体和关系的关联信息来提高模型性能[7]。Wei 等人[8]用多任务学习的方式,让实体和关系共享一个编码器(Encoder),再由两个解码器(Decoder)分别解码得到实体和关系。Zheng 等人[9]将实体关系联合抽取任务转换为序列标注任务,使用 BIO(B-begin,I-inside,O-outside)序列标注的方法对三元组进行整体建模,通过端到端模型对句子进行编码后预测每个单词所属标签类别,将临近的头尾实体及关系合并为需要抽取的三元组,取得了不错效果。但这种序列标注的方法无法处理实体嵌套的问题。Li 等人[10]则是将联合关系抽取任务转换成多轮问答任务,通过在输入中加入问题来引入实体关系信息,再由模型预测出答案,即文本中能够作为答案片段的起始结束位置。Cui 等人[11]采用序列到序列(Sequence-to-Sequence,Seq2Seq)模型,将实体关系标记作为单词加入词典,通过编码器对文本进行编码,再由解码器直接解码输出带有实体关系标记的三元组文本序列。

该文提出一种基于 Seq2Seq 的融合知识的多源关系抽取方法,通过引入规则抽取的三元组作为额外提示信息,帮助模型从文本中抽取出更准确全面的实体关系三元组。该方法能够有效利用规则知识,对于领域的迁移性较强,对于缺乏标注数据的领域,只需要利用简单的领域知识来制定一些典型规则进行抽取,抽取结果即可作为额外知识辅助模型,从而达到比直接使用 Seq2Seq 模型和单靠规则抽取都要好的效果,减少了人工定制复杂规则的成本。

2 关系抽取方法

2.1 问题描述

采用类似 Cui 等人[11]的方法,输入为一段文本和使用规则抽取出的关系元组文本,关系元组通过表示实体和关系的特殊字符进行标记后拼接成字符串,比如“_{古 龙} <rel> 本 名 </rel> <obj> 熊 耀 华 </obj>”。模型通过对文本和关系元组信息进行预测得到的所有可能三元组,并以相同方式将所有预测关系三元组进行拼接作为一个字符串输出。

2.2 整体架构

该文所面对的从网页文本中抽取所有可能实体关系这一任务的整体架构如图 1 所示。输入为来自网页的原始文本,通过预处理操作去除非正文等噪声信息,得到纯文本后使用基于领域知识制定的简单规则对文本进行关系抽取,得到后续作为提示信息的三元组,并用特殊字符标识拼接为字符序列。然后用编码器-解码器模型对文本和根据规则抽取得到的三元组字符串进行预测,输出所有可能关系元组构成的字符序列。

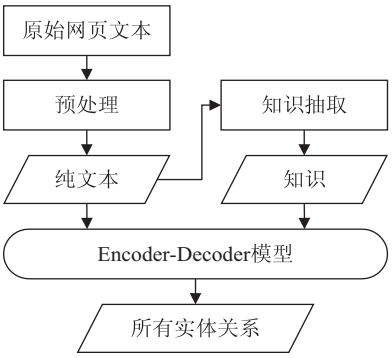

图 1 多源关系抽取方法整体架构

2.3 外部知识抽取

对于百科类的文本使用 Infobox 等结构化文本匹配和基于依存句法分析的模式匹配规则进行抽取。利用哈工大语言技术平台(Language Technology Platform,LTP)对文本进行依存句法分析,然后选取 Jia 等人[12]提出的依存语义范式进行关系抽取。对于类似企业标书等特殊文体的文本,则依据领域知识采用正则表达式、关键字匹配以及表格解析等方法进行抽取。将得到的实体对及实体间关系用特殊字符标记后组成一个形如“_{古 龙} <rel> 本 名 </rel> <obj> 熊 耀 华 </obj>”的字符串。其中“”中间的为头实体,“<obj> </obj>”中间的是尾实体,“<rel> </rel>”之间的是头尾实体间关系。

该文将上述方法抽取得到的实体关系元组信息作为每个句子的局部知识,利用知识编码器学习背后的模式、规则等全局知识来指导模型抽取,从而提高实体关系联合抽取的准确率。

2.4 模型结构

基础模型选用的是经典的 Seq2Seq 模型 Transformer[13],在其原有架构上进行改动,以引入额外知识。在其原有编码器的基础上再增加一个知识编码器,同样使用自注意力机制对引入的知识进行编码,从而获得文本中的局部知识信息。将文本编码和知识编码一同送入解码器中。在解码器增加一个解码知识的注意力层,用于解码出更类似引入知识的三元组,从而达到对知识的学习。模型结构如图 2 所示。

2.4.1 知识提取器

面对特定领域的具体任务,首先利用相关领域知识构建一个知识提取器,作为该领域全局知识。知识提取器通过使用规则从文本中抽取实体关系元组以及根据表格信息匹配对应句子等方法获取和该句子有关

的局部知识,作为输入模型的额外知识信息。

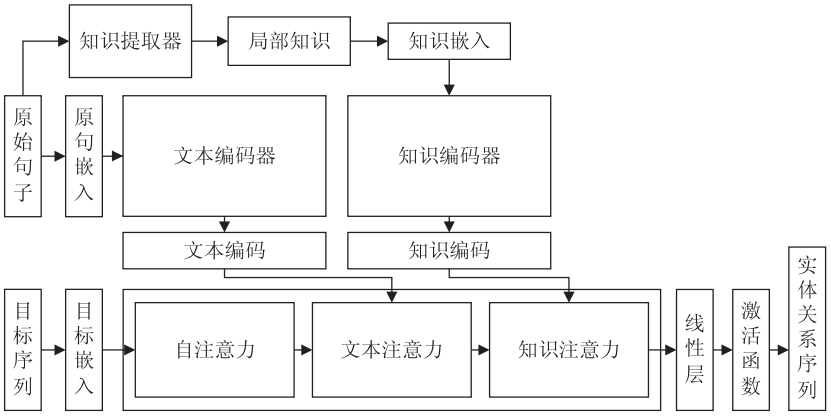

图 2　多源关系抽取模型结构

2.4.2　文本编码器

输入为句子 $X=(x_1,x_2,\cdots,x_n)$,通过词嵌入和位置编码操作后得到句子嵌入 $\boldsymbol{E}_X=(\boldsymbol{e}_{x1},\boldsymbol{e}_{x2},\cdots,\boldsymbol{e}_{xn})$。输入文本编码器后得到输入句子的上下文向量 $\boldsymbol{H}_X=(\boldsymbol{h}_{x1},\boldsymbol{h}_{x2},\cdots,\boldsymbol{h}_{xn})$

$$\boldsymbol{H}_X=\text{TextEncoder}(\boldsymbol{E}_X) \tag{1}$$

2.4.3　知识编码器

对知识提取器得到的和句子相关的知识进行特殊标记处理拼接后形成知识序列 $K=(k_1,k_2,\cdots,k_m)$,通过词嵌入和位置编码操作得到知识嵌入 $\boldsymbol{E}_K=(\boldsymbol{e}_{k1},\boldsymbol{e}_{k2},\cdots,\boldsymbol{e}_{km})$。输入知识编码器后得到知识的上下文向量 $\boldsymbol{H}_K=(\boldsymbol{h}_{k1},\boldsymbol{h}_{k2},\cdots,\boldsymbol{h}_{km})$

$$\boldsymbol{H}_K=\text{KnowledgeEncoder}(\boldsymbol{E}_K) \tag{2}$$

2.4.4　解码器

在解码阶段首先将目标三元组序列 $T=(t_1,t_2,\cdots,t_s)$ 作为解码器的输入,同样经过词嵌入和位置编码后得到目标元组序列嵌入 $\boldsymbol{E}_T=(\boldsymbol{e}_{t1},\boldsymbol{e}_{t2},\cdots,\boldsymbol{e}_{ts})$。在注意力层将 $\boldsymbol{E}_T$ 经过自注意力机制得到的目标序列隐层状态 $\boldsymbol{H}_T$ 先后与 $\boldsymbol{H}_X$ 和 $\boldsymbol{H}_K$ 计算交叉注意力,得到融合知识后的解码器输出的隐层状态 $\boldsymbol{H}_O$。

$$\boldsymbol{H}_T=\text{SelfAttention}(\boldsymbol{E}_T) \tag{3}$$

$$\boldsymbol{H}_{TX}=\text{CrossAttention}(\boldsymbol{H}_T,\boldsymbol{H}_X) \tag{4}$$

$$\boldsymbol{H}_O=\text{CrossAttention}(\boldsymbol{H}_K,\boldsymbol{H}_{TX}) \tag{5}$$

其中,$\boldsymbol{H}_{TX}$ 为解码器输入的上下文向量和编码器输入的三元组上下文向量做交叉注意力所得到的包含目标三元组和输入三元组注意力的上下文向量。

将 $\boldsymbol{H}_O$ 经过线形层和 softmax 计算得到输出单词的概率 P:

$$P=\text{softmax}(\boldsymbol{W}_l\boldsymbol{H}_O) \tag{6}$$

其中,$\boldsymbol{W}_l$ 为线性层的权重矩阵。

最后根据词表输出生成的序列 $Y=(y_1,y_2,\cdots,y_n)$。

2.5　训练方式

首先使用纯文本和所有抽取出的实体关系三元组作为编码器和解码器的输入单独训练一个 Transformer 模型并在训练好之后将各层参数固定。再用纯文本和部分抽取实体作为模型编码器输入,所有实体关系三元组作为解码器输入进行训练,训练过程中仅更新知识编码器以及目标三元组和知识交叉注意力模块的参数。

2.6　目标函数

目标函数如公式(7)所示,根据给定输入句子 X、外部知识 K 以及用标注数据训练得到的 Transformer 参数 θ_X 来生成目标序列 Y,采用最大似然估计,提升真实样本 Y 的似然概率。

$$J(D,K;\theta_x,\theta_k)=\sum_{\substack{(X,Y)\in D\\K_x\in K}}\log(Y\mid X,K_x;\theta_x,\theta_k)=\sum_{\substack{(X,Y)\in D\\K_x\in K}}\sum_{i=1}^{M}\log(y_i\mid y_{<i},X,K_x;\theta_x,\theta_k) \tag{7}$$

其中,θ_K 为知识编码器和知识交叉注意力层的参数。θ_x 为其余参数,$D=\{(X,Y)\}$ 表示监督数据集,其中 X 为句子序列,Y 为实体-关系序列,K 为全局知识图谱,K_x 为与句子相关的局部知识图谱。M 为实体-关系序列 Y 的字符个数。$y_{<i}$ 表示在预测第 i 个字符 y_i 之前已经预测出来的字符序列。

3　实验部分

3.1　数据集

Seq2Seq 模型的一个优点是比较灵活,能够通过对生成序列进行特殊构造来完成多种不同的任务。因此,该文分别在开放域和限定域两个数据集上进行实验。

实验所使用的两个数据集,一个是通用百科领域的数据集 SpanSAOKE,由 Lyu 等[14]制作,处理后包含 26 481 个中文句子以及 53 774 个实体关系三元组。其数据来源为百度发布的符号辅助开放知识表达(Symbolic Aided Open Knowledge Expression,

SAOKE)[15]数据集。该数据集是中文开放域信息抽取的大规模句子级数据集,其中每个句子都是人工标记的,并采用统一的知识表示格式来表达句子中所包含的事实。

另一个是采招网(https://www.bidcenter.com.cn/)上的标书网页文本共获取了 20 951 篇,经过去除 HTML 标签、JavaScript、CSS 等代码文本以及特殊符号等数据清洗处理后得到仅包含正文的标书文本。根据需求及对标书格式的认知制定一系列规则对标书内容进行抽取,实体类别包括招标方名称、供应商名称、招标代理机构名称、预算金额、产品名称等 24 种类别。按模型设定的 256 最大文本长度进行切分后得到总计 71 695 份原文本和事实元组的句对。将数据集划分为训练集、验证集和测试集,见表 1。

表 1 数据集统计信息

数据集	训练集	验证集	测试集	关系数
SpanSAOKE	21 183	2 647	2 651	—
Bid-Docs	64 171	7 182	342	24

其中,SpanSAOKE 为开放域关系抽取数据集,不限制关系类别,关系均从原文中抽取得到。Bid-Docs 为爬取的标书文本数据集。训练集和验证集为通过复杂的网页爬取规则进行自动化标注,测试集为自动标注后进行人工校验得到。由于该数据集比较特殊,以标书本身为头实体,只抽取关系和对应尾实体,因此标注时使用特殊字符“<key> </key>”标注关系,“<val> </val>”标注尾实体。

3.2 评价指标

实验使用 F1 值作为评价指标并采用涂飞明等人[16]在实验中所用的两种计算方式,一个是完全匹配(Exact Match,EM),对于抽取出的关系三元组,只有当预测的头实体、尾实体及实体间关系完全和标准答案相同时才算是一个正确的抽取结果。另一个是最长公共子串(Longest Common Substring,LCS),下面为具体计算方法:

考虑到模型生成多个三元组的数量和顺序不固定,因此先对答案和预测结果的三元组序列按三元组拆分后进行排序。对于答案中的每一个三元组,使用 LCS 长度在预测结果中找到其关联性最高的三元组,找到的所有三元组按答案顺序拼接,预测中多余的三元组直接按原顺序拼接在后面,形成和答案最相近的序列。然后以新得到的序列和标准答案三元组序列求出二者的最长公共子串 LCS,然后用 LCS 的长度作为预测正确的序列长度,分别与预测结果序列长度和标准答案序列长度计算精确率 P 和召回率 R,最后利用公式(8)计算 F1 值。这两种计算方法分别反映了模型预测结果的精确匹配度和模糊匹配度。在计算 LCS 时去掉了“<sub>”等用于表示头尾实体及关系的特殊符号。

$$\mathrm{F1} = \frac{2PR}{P + R} \tag{8}$$

使用 EM 指标时的精确率 P 和召回率 R 的计算公式分别为:

$$P = \frac{T_C}{T_P} \tag{9}$$

$$R = \frac{T_C}{T_G} \tag{10}$$

其中,T_C 为预测正确的三元组数量,T_P 为预测三元组总数量,T_G 为标准答案三元组总数量。

使用 LCS 指标时的精确率 P 和召回率 R 的计算公式分别为:

$$P = \frac{L_{\mathrm{LCS}}}{L_P} \tag{11}$$

$$R = \frac{L_{\mathrm{LCS}}}{L_G} \tag{12}$$

其中,L_{LCS} 为预测结果序列和标准答案序列的 LCS 的长度,L_P 为预测序列长度,L_G 为标准答案序列长度。

3.3 实验环境及参数设置

操作系统是 Ubuntu16.04.7,编程语言为 Python3.7,深度学习框架为 Pytorch1.8.1,句子最大长度设定为 256,知识编码器以及知识交叉注意力层和 Transformer 的编码器及解码器一样为 6 层,多头注意力的头数为 8 个。

3.4 实验结果

融合知识的关系抽取方法在 SpanSAOKE 和从网页获取的标书文本两个数据集上的实验结果分别如表 2 和表 3 所示,其中 Transformer 为基于注意力机制的 Seq2Seq 模型,使用自注意力机制的编码器对输入文本进行编码,再通过自回归的解码器解码出目标序列。MGD-GNN[14]是一个基于字符级的流水线式开放域关系抽取模型,通过构建单词和字符之间的依赖关系图来引入依存句法信息以及分词信息,利用图神经网络对其进行编码得到最终每个字符的表示,然后分别预测每个连续字符子序列为关系的概率,最后为每个得到的关系预测其对应的头尾实体。KGMS 为文中模型。由于标书数据集比较特殊,只抽取关系和尾实体的二元组,因此现有关系抽取方法不太适用,故只和未融合知识的 Transformer 进行比较。

在 SpanSAOKE 数据集上,相比之下 MGD-GNN 因为是对句子的每个连续字符子序列进行预测,所以抽取到了最多的关系元组,具有最高的召回率。但其预测正确的元组数量没有 KGMS 所抽取的多,并且对于没有完全匹配的关系元组,KGMS 也得到了更接近

答案的结果。由此可见,KGMS 在抽取时参考了输入元组所包含的规则信息,使得抽取更具有针对性,从而提高了完全匹配的精确率和召回率。

表 2　SpanSAOKE 实验结果

模型		实体			关系			实体关系		
		P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
EM	Transformer	0.50	0.47	0.48	0.41	0.9	0.40	0.24	0.22	0.23
	MGD-GNN	0.55	0.59	0.56	0.48	0.54	0.51	0.31	0.35	0.32
	KGMS	0.70	0.52	0.60	0.65	0.47	0.55	0.38	0.30	0.34
LCS	Transformer	0.58	0.64	0.57	0.52	0.56	0.48	0.71	0.72	0.67
	MGD-GNN	0.57	0.69	0.58	0.55	0.62	0.52	0.64	0.72	0.63
	KGMS	0.68	0.61	0.62	0.66	0.59	0.60	0.80	0.69	0.70

表 3　标书文本实验结果

模型		实体			关系			实体关系		
		P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
EM	Transformer	0.03	0.01	0.02	0.53	0.28	0.37	0.03	0.01	0.02
	KGMS	0.09	0.06	0.07	0.58	0.45	0.50	0.06	0.04	0.05
LCS	Transformer	0.09	0.07	0.07	0.38	0.28	0.28	0.33	0.20	0.21
	KGMS	0.19	0.14	0.15	0.49	0.42	0.43	0.39	0.27	0.30

在标书文本上,KGMS 在引入知识信息后性能有所提升,但整体表现都不好,分析可能是因为网页标书格式比较多样,且有一些表格,导致爬取后的数据十分杂乱,且难以通过预处理进行修整。由于标书文本形式比较特殊,默认将标书题目作为所有要抽取的关系元组的头实体,在抽取时仅抽取定义的关系及对应尾实体。很多时候定义的关系并未出现在原文中,需要模型根据上下文来判断实体所属类型,而上下文信息又往往比较零碎,严重影响了模型的特征提取。

3.5　实例分析

Text:网游公司的地推人员通常分为两类:一类是正式招聘,多为大学生;

Gold:<地推人员,分为,两类><一类,是,正式招聘><地推人员,多为,大学生>

Pattern Matching:<地推人员,多为,大学生><网游公司地推人员,分为,两类>

Transformer:<网游公司的地推人员,通常分为,两类><一类,是,正式招聘><多,为,大学生>

MGD-GNN:<网游公司的地推人员,通常分为,两类><一类,是,正式招聘><网游公司的地推人员,多为,大学生>

KGMS:<地推人员,多为,大学生><地推人员,分为,两类><网游公司推司,分为,两类>

如上所示,相比之下 Transformer 和 MGD-GNN 所抽取的实体都包含了修饰语,而从标准答案来看显然是不希望修饰语出现的。Pattern Matching 为通过规则获取的作为额外知识的三元组。其中包含了有修饰语和无修饰语的实体,使得 KGMS 在引入知识后尝试将两种类型都抽取出来。但由于 Pattern Matching 没有关于“一类是正式招聘”,从而导致 KGMS 忽略了这部分。由此可说明引入的知识对实体关系抽取有比较大的影响,可以通过对额外知识的设计,引入更多更符合需求的知识来对模型的抽取行为进行引导,从而提升模型的性能。

4　结束语

该文提出了一种融合外部知识的实体关系联合抽取方法,采用多编码和知识注意力机制,将句法结构和结构化信息等外部知识融入编码解码框架的模型来生成实体和关系序列,从而提高模型的准确率。实验结果显示,在开放域和限定域的数据集上的表现均有所提升,可以提升抽取准确率。对于一些标注数据稀缺场景是有效的提升方法,同时规则的制定可以让模型更好地具有领域适应性。在今后的工作中将使用预训练模型作为基础模型,并构建更完善的数据集来探究该方法。

参考文献:

[1]　李冬梅,张　扬,李东远,等. 实体关系抽取方法研究综述[J]. 计算机研究与发展,2020,57(7):1424-1448.

[2]　刘　辉,江千军,桂前进,等. 实体关系抽取技术研究进展综述[J]. 计算机应用研究,2020,37(S7):1-5.

[3] MILLER S, FOX H, RAMSHAW L, et al. A novel use of statistical parsing to extract information from text[C]//the 1st meeting of the north American chapter of the association for computational linguistics. Seattle: ACM, 2000: 226-233.

[4] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. The Journal of Machine Learning Research, 2003, 3: 1083-1106.

[5] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing of the AFNLP. Suntec: Association for Computational Linguistics, 2009: 1003-1011.

[6] ZHONG Z, CHEN D. A frustratingly easy approach for entity and relation extraction[C]//North American chapter of the association for computational linguistics. [s. l.]: Association for Computational Linguistics, 2021: 50-61.

[7] 李代祎,李忠良,严 丽. 一种面向中文的实体关系联合抽取方法研究[J/OL]. 小型微型计算机系统: 1-9[2022-10-07]. http://kns. cnki. net/kcms/detail/21. 1106. TP. 20220727. 1525. 004. html.

[8] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]//58th annual meeting of the association for computational linguistics. [s. l.]: Association for Computational Linguistics, 2020: 1476-1488.

[9] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv: 1706. 05075. 2017.

[10] LI X Y, YIN F, SUN Z J, et al. Entity-relation extraction as multi-turn question answering[J]. arXiv: 1905. 05529, 2019.

[11] CUI L, WEI F R, ZHOU M. Neural open information extraction[C]//56th annual meeting of the association for computational linguistics (volume 2: short papers). Melbourne: Association for Computational Linguistics, 2018: 407-413.

[12] JIA S B, LI M Z, XIANG Y, et al. Chinese open relation extraction and knowledge base establishment[J]. ACM Transactions on Asian & Low Resource Language Information Processing, 2018, 17(3): 1-22.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv: 1706. 03762, 2017.

[14] LYU Z H, SHI K J, LI X, et al. Multi-grained dependency graph neural network for Chinese open information extraction[C]//25th Pacific-Asia conference on knowledge discovery and data mining. [s. l.]: Springer, 2021: 155-167.

[15] SUM M M, LI X, WANG X, et al. Logician: a unified end-to-end neural approach for open-domain information extraction[C]//The eleventh ACM international conference on web search and data mining. New York: Association for Computing Machinery, 2018: 556-564.

[16] 涂飞明,刘茂福,夏 旭,等. 基于 BERT 的阅读理解式标书文本信息抽取方法[J]. 武汉大学学报: 理学版, 2022, 68(3): 311-316.

(上接第 123 页)

[18] 陈 瀛,生佳根. 基于 LDA 和 Word2vec 的微博标签生成算法[J]. 计算机与现代化, 2021(12): 37-42.

[19] 王恩慧. 社交网络中的观点挖掘与情感动因分析方法研究[D]. 北京: 北京交通大学, 2021.

[20] LIANG X, LIU P, WANG Z. Hotel selection utilizing online reviews: a novel decision support model based on sentiment analysis and DL-VIKOR method[J]. Technological and Economic Development of Economy, 2019, 25(6): 1139-1161.

[21] ZHANG J, ZHANG A, LIU D, et al. Customer preferences extraction for air purifiers based on fine-grained sentiment analysis of online reviews[J]. Knowledge-Based Systems, 2021, 228: 107259.

[22] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301. 3781, 2013.

[23] LI G, LAW R, VU H Q, et al. Identifying emerging hotel preferences using emerging pattern mining technique[J]. Tourism Management, 2015, 46: 311-321.

[24] XIAO S, WEI C P, DONG M. Crowd intelligence: analyzing online product reviews for preference measurement[J]. Information & Management, 2016, 53(2): 169-182.

[25] ZHANG W, XU H, WAN W. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis[J]. Expert Systems with Applications, 2012, 39(11): 10283-10291.

[26] 张 林. 基于 web 的定制产品用户评论情感分析系统[D]. 大连: 大连理工大学, 2021.

[27] SCHÜTZE H, MANNING C D, RAGHAVAN P. Introduction toinformation retrieval[M]. Cambridge: Cambridge University Press, 2008.