

基于跨度和语义特征的实体关系抽取模型

冯 萍^{1,2}, 杨芷茜¹, 王韧杰¹, 冯师语¹, 吴 航¹, 孙 宇³

(1. 长春大学 计算机科学技术学院, 长春 130022; 2. 长春大学 残障人士智能康复与无障碍教育部重点实验室, 长春 130022; 3. 长春大学 特殊教育学院, 长春 130022)

摘 要: 针对传统实体关系抽取方法依赖距离度量或简单跨度识别, 难以捕捉实体间潜在关系, 且模型计算复杂度高、易引发误差传递的问题, 提出了一种基于跨度和语义特征的端到端实体关系抽取模型。首先, 将文本向量随机分割成跨度序列, 以便模型能够学习更广泛的语义特征信息。其次, 对语义关系进行判定, 以筛选出候选关系子集, 从而减少信息冗余。最后, 将候选关系转换为包含重要关系语义的关系-跨度组合, 并利用 Transformer 解码器实现实体关系联合抽取。实验结果表明, 相较于其他基线模型, 该模型在 NYT 数据集和 WebNLG 数据集上的 F_1 值均有显著提升, 证明了其有效性。

关键词: 实体关系抽取; 跨度; 语义特征; 上下文信息

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2025)12-4045-07

DOI: 10.13229/j.cnki.jdxbgxb.20240474

Entity relationship extraction method based on span and semantic features

FENG Ping^{1,2}, YANG Zi-qian¹, WANG Ren-jie¹, FENG Shi-yu¹, WU Hang¹, SUN Yu³

(1. College of Computer Science and Technology, Changchun University, Changchun 130022, China; 2. Ministry of Education Key Laboratory of Intelligent Rehabilitation and Barrier-free Access for the Disabled, Changchun University, Changchun 130022, China; 3. College of Special Education, Changchun University, Changchun 130022, China)

Abstract: A end-to-end entity relationship extraction model based on span and semantic features was proposed to address the problems of traditional entity relationship extraction methods relying on distance measurement or simple span recognition, which make it difficult to capture potential relationships between entities, and the high computational complexity and error propagation of the model. Firstly, the text vectors were randomly segmented into span sequences so that the model can learn a wider range of semantic feature information. Secondly, semantic relationships were judged to screen out subsets of candidate relationships, thus reducing information redundancy. Finally, the candidate relationships were transformed into relationship-span combinations containing important relationship semantics, and the Transformer decoder was used to achieve the joint extraction of entity relationships. The experimental

收稿日期: 2024-04-30.

基金项目: 吉林省科技厅科技发展计划项目(2023JB405L07); 国家自然科学基金面上项目(62377006); 吉林省发改委计划项目(2019C048-6).

作者简介: 冯萍(1977-), 女, 副教授, 博士研究生. 研究方向: 人工智能、知识图谱、康复工程.

E-mail: fengping@ccu.edu.cn

results show that the F_1 value of this model is significantly improved in the NYT and WebNLG datasets compared to other baseline models, proving its effectiveness.

Key words: entity relationship extraction; span; semantic feature; context information

0 引言

实体关系抽取是将自然语言转化为结构化数据的重要任务,其目标是从非结构化文本中识别实体及其之间的关系,生成实体关系三元组,为各种自然语言处理任务提供支持并奠定基础。^[1]

实体关系抽取目前主要可以分为流水线方法和联合抽取方法两种方法。流水线抽取方法通常包括命名实体识别^[2-5]和关系抽取^[6,7]两个子任务。虽然该方法可以分割成多个任务模块分别进行实体识别和关系抽取,但各任务之间经常存在信息冗余问题。因此,如何更好地减少信息冗余和误差传递是实体关系抽取的研究重点。

近年来,研究学者提出了一系列联合抽取的方法。联合抽取将实体抽取和关系抽取两个子任务进行联合,能够充分利用实体和关系之间的交互信息,从而提供更全面的文本理解^[8],解决模块之间、步骤之间存在的级联冗余问题,减少误差传递。Lin 等^[9]使用卷积神经网络嵌入句子的语义,通过引入选择性注意机制,在关系抽取过程中对实例进行动态加权,从而更充分地利用语义信息。但该方法存在数据不均衡和噪声数据问题,影响模型的可用性。Kalpit 等^[10]提出了一种用于联合实体提及检测和关系提取的跨级端到端模型。相较于传统 token 级模型,该模型可充分挖掘跨度专属特征,支持重叠实体提及的识别,且无需借助顺序解码机制。然而,该模型存在计算复杂度过高的缺陷,当处理长文本时,极易生成大量负样本,影响模型的整体性能。Eberts 等^[11]提出了跨度感知联合抽取模型(SPERT),同时使用了负采样的方式提高模型性能,并通过最大池化的方式,提取关系中实体对之间的文本特征。与传统基于序列标注的方法相比,SPERT 可以处理更复杂的实体和关系结构,具有更好的泛化能力。

在 SPERT 模型的基础上,Wan 等^[12]提出了跨度-掩码联合抽取模型(SMAN),该模型使用了完型机制(Cloze mechanism, CM)同时提取上下文和 span 位置的信息,并在关系抽取阶段对 span 和标签进行建模,但需要大量的文本数据和对应的标注用于创建 cloze 任务。

传统的抽取模型通常依赖距离度量或简单的跨度识别方法,这些方法在处理关系时未能充分利用深层次的语义信息。它们更多从符号角度处理数据,因此容易忽略关系本身的含义,以及上下文语义与关系之间的内在联系。

为有效解决以上问题,本文提出一种基于跨度和语义特征的端到端的实体关系抽取模型。首先,采用统一跨度的方法处理不同长度的实体,避免将其切分成单独的词语;其次,利用多头注意力机制捕获关系的语义特征,并筛选候选关系子集;最后,在捕获关系语义的同时结合实体的语义信息提取头尾实体,并通过共享参数实现实体识别模块与关系抽取模块之间的交互。实验结果表明,该模型将语义信息与实体关系抽取相结合,实现了更综合、更高效的抽取过程。

1 基于跨度和语义特征的实体关系抽取模型

1.1 模型结构

模型主要由文本嵌入模块(Text embedding module, TEM)、基于上下文的语义关系判定模块(Semantic relationship determination module, SRDM)和实体关系联合抽取模块(Entity relationship joint extraction module, ERJEM)3 个部分组成,总体结构如图 1 所示。

在文本嵌入模块,首先将文本输入来自变换器的双向编码器表示(Bidirectional encoder representations from transformers, BERT)^[13]嵌入层,使用 BERT 预训练语言模型将文本转换为向量,并将向量随机分割成跨度序列。在语义关系判定模块中,使用多头注意力机制筛选最相关的关系子集,过滤不相关的关系,提高关系细粒度识别的精确度,以便对跨度和关系进行联合判定。在实体关系联合抽取模块中,将候选关系转换为包含重要关系语义的关系-跨度组合,并使用 Transformer 解码器完成三元组抽取任务。

1.2 文本嵌入模块

近年来,以 BERT 为代表的预训练模型在自然语言处理(Natural language processing, NLP)的

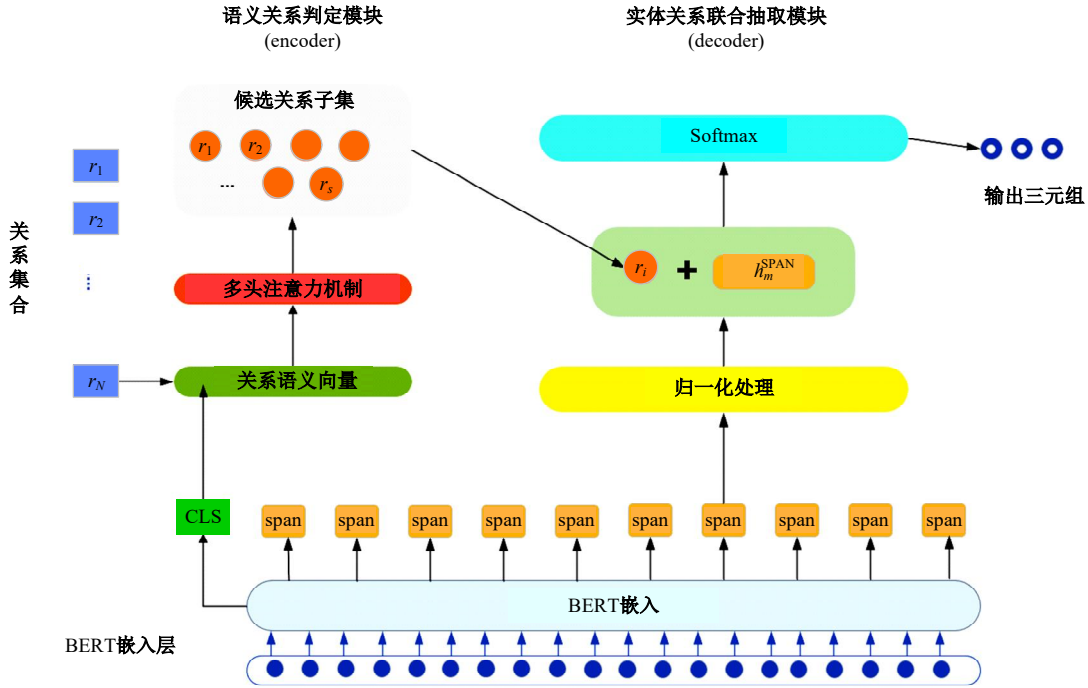


图1 Semantic-SPAN 实体关系抽取模型

Fig. 1 Semantic-SPAN entity relation extraction model

各项任务中取得了较先进的效果,也被成功应用到实体识别、关系抽取任务中,因此本文选用BERT作为文本嵌入工具。定义输入文本为 $X = \{x_1, x_2, \dots, x_n\}$, x_n 表示文本中的字符, n 表示文本序列的长度。将字符级分割后的文本输入BERT模型。BERT预训练模型包含12个隐藏层,每个隐藏层的大小为768。每个字符被映射为对应的向量表示,如下所示:

$$H = \text{BERT}(X) = h_1, h_2, \dots, h_n \quad (1)$$

式中: H 为经过BERT模型处理后某一跨度内所有字符的向量表示集合; h_t 为第 t 个字符的向量表示。

引入跨度表示,对于特定的跨度,本文关注由特殊标记[START]和[END]所界定的区域,[START]和[END]分别表示跨度的开始位置和结束位置。设[START]和[END]的向量分别表示为 h_{START} 和 h_{END} ,跨度语义的向量表示记为 h^{SPAN} 。同时生成上下文标记[CLS](见图1),以表示整个句子(或上下文)的语义。

1.3 语义关系判定模块

传统的实体关系抽取一般基于距离度量或简单跨度识别,未能充分挖掘实体之间的潜在关系。例如“Xiao Ming's father is Wang Jian-jun.”中存在“father”这一关系,由于语义的多样性还会出现例如“Wang Jian-jun's son is named Xiao Ming.”

的情况,这类隐含关系在文本语义层面十分明显。

本文提出了一种联合语义关系识别与关系过滤的语义关系判定方法。如图2所示,该方法对数据集中全部关系集合 $R = \{r_1, r_2, \dots, r_N\}$ 进行遍历,分别计算各关系与当前待抽取语句在语义层面的相似度,并以二分类的方式判定关系为相关或不相关,这一过程极大地提高了后续关系细粒度识别的精确度。

首先,获取[CLS]位置的向量信息并记为 $x_{\text{BERT}}^{\text{CLS}}$,该向量包含了当前输入语句 X 的语义信息;接着,以同样的方式获取关系的语义向量记为 $r_{\text{BERT}}^{\text{CLS}}$;最后,通过维基百科对关系的语句中的关系词条进行查询并获取文本描述语句。语义关系判定概率预测 p_r 如下所示:

$$p_r = \text{Softmax} \left[W_r \text{ATT} \left(x_{\text{BERT}}^{\text{CLS}}, r_{\text{BERT}}^{\text{CLS}} \right) + b_r \right] \quad (2)$$

式中:ATT($\cdot$)为关系语义注意力运算函数; W_r 为可训练权重; b_r 为偏置。

此处采用缩放点积注意力机制,如下所示:

$$\text{Attention}(Q, K, V) = \text{Soft max} \left(\frac{Q \cdot K^T}{\sqrt{d_k}} \right) \cdot V \quad (3)$$

式中: Q 为索引值; K 为关键字索引值; V 为关键字值; $\sqrt{d_k}$ 为矩阵 K 的维度。

同时,设置关系过滤阈值,若 p_r 的值大于关系过滤阈值的值,则认为是有效的关系类型 p_r ,否则

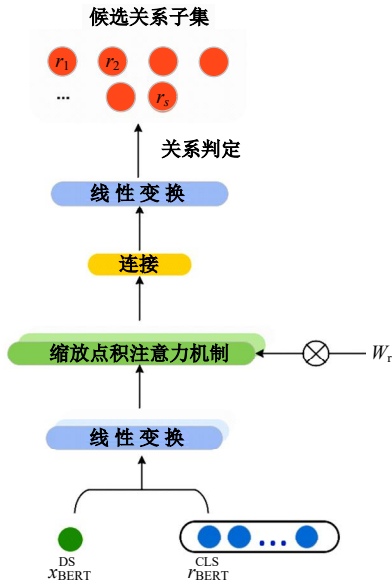


图2 语义关系判定模块结构

Fig. 2 Semantic relationship determines the module structure

认为是无效的关系类型。

通过上述计算后,模型可以过滤掉不相关的关系,获得候选关系子集 $R' = \{r_1, r_2, \dots, r_s\}$ 。

1.4 实体关系联合抽取模块

在完成语义关系判定后,将候选关系与跨度进行组合,完成基于跨度与语义特征的实体关系联合抽取,模块结构如图3所示。

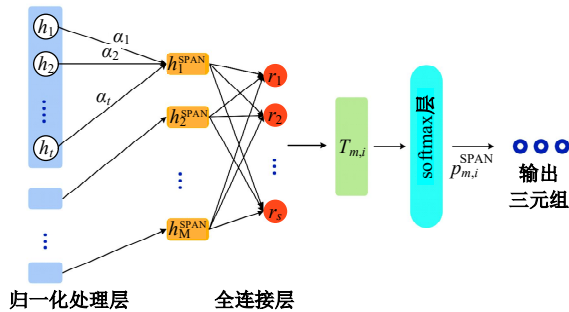


图3 联合抽取模块结构

Fig. 3 Structure of joint extraction module

由1.2节可知,跨度语义向量表示为 h^{SPAN} ,如下所示:

$$h^{\text{SPAN}} = \sum_{i=1}^n \alpha_i h_i \quad (4)$$

α_i 定义为:

$$\alpha_i = \frac{\exp[\text{score}(h_i)]}{\sum_{j=1}^n \exp[\text{score}(h_j)]} \quad (5)$$

式中: $\text{score}(h_i)$ 为得分函数,用于评估字符 x_i 的重要性; α_i 为归一化的注意力权重。

得分函数 $\text{score}(q, k)$ 如下所示:

$$\text{score}(q, k) = \frac{q^T k}{\sqrt{d_k}} \quad (6)$$

式中:输入为查询 $q \in \mathbb{R}^q$ 和键 $k \in \mathbb{R}^k$ 。

通过结合字符级的粒度和BERT的深层语义理解能力,有效捕捉跨度 h^{SPAN} 内的细粒度语义信息,为复杂的三元组抽取任务提供更强大的语义支持。跨度语义向量集合表示为 $H^{\text{SPAN}} = \{h_1^{\text{SPAN}}, h_2^{\text{SPAN}}, \dots, h_M^{\text{SPAN}}\}$ 。

每个关系都需要与可能的跨度相结合进行概率计算,因此输入关系跨度组表示为 $T_{m,i}$,如下所示:

$$T_{m,i} = (h_m^{\text{SPAN}}, r_i) \quad (7)$$

式中: r_i 为第 i 个候选关系; h_m^{SPAN} 为第 m 个跨度语义向量。

接下来的目标是提取相应的实体对,即判断 $T_{m,i}$ 的概率大小,使用 softmax 激活函数预测该跨度与候选关系匹配实体对的概率 $p_{m,i}^{\text{SPAN}}$,如下所示:

$$p_{m,i}^{\text{SPAN}} = \text{softmax}(T_{m,i}) \quad (8)$$

1.5 损失函数

考虑到模型中总共有两个任务:关系判定和联合提取。因此在优化过程中,本文以多任务的方式联合训练模型,并共享编码器的参数。为了预测实体对,按它们在文本中的顺序进行排序,并采用交叉熵损失作为三元组提取的损失函数 L ,如下所示:

$$L = - \sum_{i=1}^S \sum_{m=1}^M \log(p_r p_{m,i}^{\text{SPAN}}) \quad (9)$$

式中: S 为候选关系数量; M 为候选跨度数量。

2 实验和结果分析

2.1 数据集

本文在 NYT 数据集^[14]和 WebNLG 数据集^[15,16]两个基准数据集上对本文模型进行评估。NYT 数据集是一个常用于关系抽取和知识图谱构建的大型语料库,它包含丰富的真实世界语言数据,涵盖各种主题和领域的文章。WebNLG 数据集是一个常用于自然语言处理的数据集,由结构化数据及其对应的自然语言描述构成,核心用途是支撑“结构化数据到自然语言”的文本生成任务,为研究者提供了丰富的实验材料和评估基准。

因此,选取以上两个数据集用于评估本模型

在实体关系联合抽取方面的性能。

2.2 参数设置

本文模型采用 NVIDIA GeForce RTX4090 (24G) 图形处理单元 (Graphics processing unit, GPU), 硬件设备配置表及参数、超参数的设置等如表 1 所示。

表 1 实验环境配置

Table 1 Experimental environment setting

参数	值
GPU/(24 G)	4 090
内存/G	32
batch size	16
学习率	0.000 1
max span	6
epoch	100
BERT_emb	768

2.3 评价指标

本文使用的评价指标为准确率 P 、召回率 R 和 F_1 指数。 F_1 是实体关系提取中最重要的评价参数。以上评价指标的计算过程如式 (10)~(12) 所示:

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (12)$$

式中: TP 为真正例 (True positive), 即模型预测为正类且实际为正类的样本数量; FP 为假正例 (False positive), 即模型预测为正类但实际为负类的样本数量; FN 为假负例 (False negative), 即模型预测为负类但实际为正类的样本数量。

2.4 结果与分析

2.4.1 对比实验

本文将模型与以下 4 个强基线联合抽取模型进行了比较, 具体如下。

CopyRE^[17]: 提出了一种基于注意力机制的抽取模型, 在抽取实体关系时复制实体的上下文信息。

GraphRel^[9]: 提出了一种端到端抽取模型, 利用关系加权图卷积网络 (Graph convolutional network, GCN) 联合实体识别和关系抽取。

PRGC^[10]: 该模型综合利用实体间的潜在关系和全局信息, 以提升实体关系抽取的性能。

GRTE^[18]: 提出了一种基于填表的面向全局

特征的三重提取模型, 表填充机制能够有效捕捉实体与关系之间的特征信息。

各模型具体对比结果如表 2 所示, 本文模型在 WebNLG 数据集上准确率达 94.0%、召回率达 94.1%, F_1 值达 94.0%, 相较于较为主流的 CopyRE 模型和 GraphRel 模型具有一定程度的提升, 与表现相对优秀的基准模型 GRTE 对比, F_1 值提升了 0.1%, 这主要是因为语义关系判定模块对关系进行过滤, 减少了不相关关系出现而导致的冗余误差, 进而提高了关系识别效果, 同时降低了后续三元组概率计算量。其次, 引入关系语义信息与实体语义信息对关系三元组抽取具有重要意义, 这使得关系、头实体、尾实体不再相互孤立, 更容易发现三元组的隐含关系。

表 2 模型比较的实验结果

Table 2 Experimental results of model comparison

模型	NYT			WebNLG		
	P	R	F_1	P	R	F_1
CopyRE	61.0	56.6	58.7	37.7	36.4	37.1
GraphRel	63.9	60.0	61.9	44.7	41.1	42.9
PRGC	89.6	82.3	85.8	90.6	88.5	89.5
GRTE	92.9	93.1	93.0	93.7	94.2	93.9
Ours	93.1	93.8	93.4	94.0	94.1	94.0

仅对比主流 GRTE 模型与本文模型, GRTE 模型需要对整体文本进行全局计算, 计算复杂度高; 本文模型虽然使用了更多的子模块, 但其优化效率依然较高, 模型训练过程中的损失变化曲线如图 4 所示。

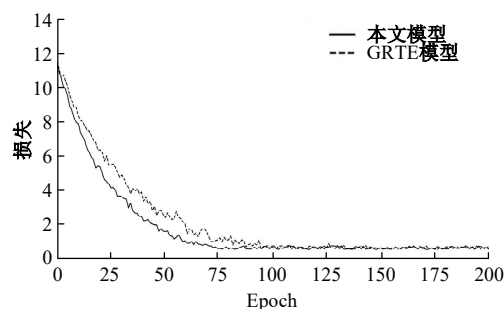


图 4 损失变化曲线 (与 GRTE 模型对比)

Fig. 4 Loss change curve (compared to the GRTE model)

2.4.2 消融实验

为进一步验证本文模型的有效性, 对实验细节展开深入分析, 重点探讨参数设置与子模块设计等对模型性能的影响。选取关系筛选模块的有

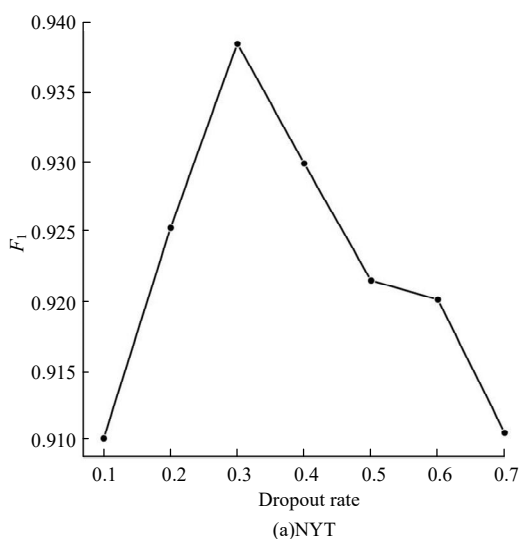
无作为变量,在上述两个数据集上进行消融实验,结果如表 3 所示。

本文模型通过语义关系判定筛选提高了识别精度,证明了语义筛选关系子集的有效性。另外,

表 3 本文模型在两个数据集上的消融实验结果

Table 3 Results of ablation experiments of the proposed model on two datasets

数据集	模式	$F_1/\%$
NYT	语义筛选关系子集	93.4
	常规	90.3
WebNLG	语义筛选关系子集	94.0
	常规	88.7



由于关系子集的筛选提高了模型的训练效率,相较于未进行关系筛选的方式,每个训练周期耗时减少了 60%。

2.4.3 调试正则化参数实验

在训练过程中,本文同时考虑 Dropout 参数对模型的影响。Dropout 随机地将网络中一些神经元的输出设置为 0,从而在每次训练迭代中随机“丢弃”一些神经元^[19]。这样有助于降低神经网络的复杂性,强制网络中的不同部分学习到更鲁棒、独立的特征,从而提高模型的泛化能力。

本文模型在不同 Dropout 参数下的 F_1 值变化曲线如图 5 所示。

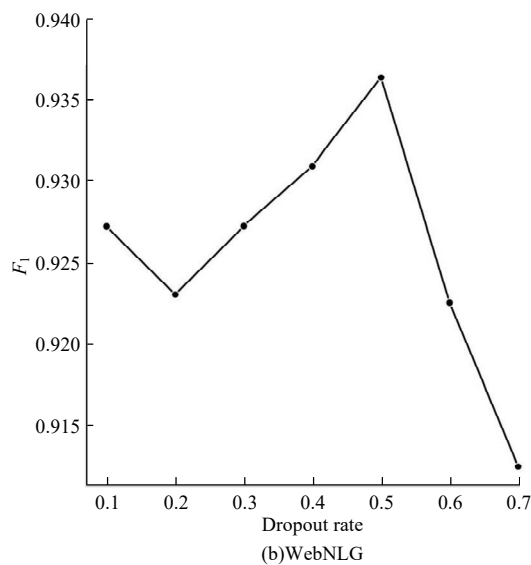


图 5 不同 Dropout 参数下 F_1 值的变化曲线

Fig. 5 Change curves of F_1 values for different Dropout parameters

由图 5 可知,在 NYT 数据集和 WebNLG 数据集上,当 Dropout 丢弃率分别设置为 0.3 和 0.5 时,本文模型能获得更优秀的 F_1 值,减少了神经网络对于特定神经元的依赖,提高了模型的泛化能力。

3 结束语

本文提出一种基于跨度和语义特征的端到端实体关系联合抽取模型,该模型充分利用实体语义和关系语义,为抽取任务提供丰富的上下文信息,并能够同时抽取实体并识别实体对之间的关系,有效解决了传统方法中模块间交互性弱、步骤级联冗余的问题。首先,跨度标签使模型能够学习到更广泛的语义信息,而非局限于独立的词语或实体,从而提升了模型对上下文的理解能力,进

一步增强了模型的泛化性能,减少了误差传递。其次,基于语义特征的关系判定模块进行关系过滤,以避免出现不相关的关系,减少信息冗余,降低了关系抽取的计算复杂度,同时提高了模型的一致性和准确性。实验结果表明,在英文关系抽取数据集 NYT 和 WebNLG 上,本文模型在 F_1 值上展现出明显优势,验证了其有效性。

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.
Li Dong-mei, Zhang Yang, Li Dong-yuan, et al. Review of studies on entity relationship extraction methods[J]. Computer Research and Development, 2020, 57(7): 1424-1448.

- [2] Yang Y, Wu Z L, Yang Y X, et al. A Survey of information extraction based on deep learning[J]. Applied Sciences, 2022, 12(19): No. 9691.
- [3] Li J, Sun A X, Han J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020: 50-70.
- [4] 王进, 蒋诗琪. 基于多路局部特征整合的嵌套命名实体识别方法[J]. 江苏大学学报:自然科学版, 2025, 46(4): 431-437.
Wang Jin, Jiang Shi-qi. Nested named entity recognition method based on multiplexed local feature integration[J]. Journal of Jiangsu University (Natural Science Edition), 2025, 46(4): 431-437.
- [5] 王进, 王猛旗, 张昕跃, 等. 基于多头注意力机制字词联合的中文命名实体识别[J]. 江苏大学学报:自然科学版, 2024, 45(1): 77-84.
Wang Jin, Wang Meng-qi, Zhang Xin-yue, et al. Chinese named entity recognition based on multi-head attention character word integration[J]. Journal of Jiangsu University (Natural Science Edition), 2024, 45(1): 77-84.
- [6] Tuo M, Yang W Z, Wei F Y, et al. A novel chinese overlapping entity relation extraction model using word-label based on cascade binary tagging[J]. Electronics, 2023, 12(4): No. 1013.
- [7] Wang P. A survey of research on deep learning entity relationship extraction[J]. Natural Language Processing and Speech Recognition, 2019, 1(1): 1-5.
- [8] Li J L, Xu Y J, Lin H Z, et al. Semantic-consistent learning for one-shot joint entity and relation extraction[J]. Applied Intelligence, 2022, 53(5): 5963-5976.
- [9] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 2124-2133.
- [10] Kalpit D, Yaser A. Span-Level Model for Relation Extraction[C]//Proceedings of 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 2019: 5308-5314.
- [11] Eberts M, Ulges A. Span-based joint entity and relation extraction with transformer pre-Training[C]//Proceedings of the 24th European Conference on Artificial Intelligence. Amsterdam: IOS Press, 2020: 2006-2013.
- [12] Wan Q, Wei L, Zhao S, et al. A span-based multi-modal attention network for joint entity-relation extraction[J]. Knowledge-Based Systems, 2023, 262: No. 110228.
- [13] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-03-15]. <https://doi.org/10.48550/arXiv.1810.04805>
- [14] Ma L, Ren H, Zhang X. Effective cascade dual-decoder model for joint entity and relation extraction[EB/OL]. [2024-03-17]. <https://arxiv.org/pdf/2106.14163v1>
- [15] Surdeanu M, Tibshirani J, Nallapati R, Manning C D. Multi-instance multi-label learning for relation extraction[C]//Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012: 455-465.
- [16] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Berlin: Springer, 2010: 148-163.
- [17] Zeng X R, Zeng D J, He S Z, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018: 506-514.
- [18] Ren F L, Zhang L H, Yin S J, et al. A novel global feature-oriented relational triple extraction model based on table filling[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 2646-2656.
- [19] Xie X Y, Xie M, Moshayedi A J, et al. A hybrid improved neural networks algorithm based on L2 and dropout regularization[J]. Mathematical Problems in Engineering, 2022: 1-19.