

面向中文短文本的实体识别与属性抽取联合模型

展辉辉 叶佩雯 李默宇

陆军工程大学军械士官学校

DOI:10.32629/acair.v4i3.21353

[摘要] 中文短文本上下文稀疏、标注噪声突出,给实体识别与属性抽取带来挑战;传统流水线方法割裂两任务语义关联,导致错误传播。为此构建面向中文短文本的联合模型:共享编码层统一捕获跨任务语义表示,边界感知模块应对实体边界模糊,实体驱动解码模块完成属性-值对抽取,单向跨模块语义传递机制将实体识别结果注入属性抽取过程,联合优化目标针对样本稀疏与噪声专项设计。实验表明联合模型多项指标优于流水线基线,各模块有效性得到验证。

[关键词] 中文短文本; 实体识别; 属性抽取; 联合模型

中图分类号: TP316.9 **文献标识码:** A

Joint Model for Entity Recognition and Attribute Extraction in Chinese Short Texts

Huihui Zhan Peiwen Ye Moyu Li

Army Engineering University School of Armament Officers

[Abstract] Chinese short texts are characterized by sparse context and significant annotation noise, posing challenges for entity recognition and attribute extraction. Traditional pipeline methods sever the semantic connections between these two tasks, leading to error propagation. To address this, a joint model for Chinese short texts is constructed: a shared encoding layer captures cross-task semantic representations; a boundary-aware module handles ambiguous entity boundaries; an entity-driven decoding module performs attribute-value pair extraction; a unidirectional cross-module semantic transfer mechanism injects entity recognition results into the attribute extraction process; and a joint optimization objective is specifically designed to mitigate sample sparsity and noise. Experimental results demonstrate that the joint model outperforms pipeline baselines across multiple metrics, validating the effectiveness of each module.

[Key words] Chinese short texts; entity recognition; attribute extraction; joint model

引言

实体识别与属性抽取在语义层面高度耦合,实体边界的准确识别直接制约属性-值对的抽取质量^[1]。中文短文本场景下,上下文窗口极度压缩,词汇歧义难以通过周边语境消解,标注样本稀缺且噪声突出,传统流水线方法独立建模两个任务,既无法利用任务间的互促信号,又放大了错误的级联效应。针对短文本场景的特殊性质,从共享编码、边界感知、实体驱动解码到联合优化目标,系统性构建模块边界清晰的联合抽取框架,以打通两个任务间的语义通路。

1 中文短文本实体识别与属性抽取的挑战与联合建模的必要性

中文短文本在社交媒体、电商评论及搜索查询等真实场景中大量存在,其上下文窗口极度压缩、词汇歧义难以通过周边语境消解、实体边界模糊且属性表达高度隐式,这些特性使得实体识别与属性抽取任务面临远比长文本更为严峻的语义理

解挑战^[2]。传统流水线方法将两任务割裂为独立子任务依次执行,实体识别阶段的边界判断误差会直接传导至属性抽取阶段形成不可逆的错误级联效应,两任务之间的语义互促信号也因任务隔离而被完全丢弃,短文本场景下标注样本稀缺与噪声比例偏高,单任务独立建模的方式无法有效应对这一数据层面的困境,模型泛化能力受到严重制约。这一系列挑战的根源在于两任务间深层语义依存关系的长期忽视,单任务建模范式在短文本场景下已难以为继,面向中文短文本的联合建模因此成为亟待解决的研究课题。

2 面向中文短文本的实体识别与属性抽取联合模型

2.1 基于预训练语言模型的跨任务共享编码层

共享编码层以中文预训练语言模型为基础骨架,将实体识别与属性抽取两项任务的输入统一映射至同一高维语义空间,从编码阶段根本上消除两任务间的表示隔阂^[3]。针对中文短文本的字符级特性,编码层采用字级别切分策略,将输入序列切分

为单字token, 最大序列长度设定为128。区别于通用预训练模型的标准层归一化设计, 编码层为实体识别与属性抽取两条任务支路分别维护独立的归一化参数, 使共享表示在进入各自任务解码器之前完成任务导向的语义校准, 隐层维度设定为768。共享编码层的参数量约占整体模型的85%, 两任务的梯度信号在反向传播过程中同步流入共享层, 通过梯度加权融合策略对两任务的梯度贡献按损失比值动态分配权重, 各任务的梯度分配系数在每个训练步依据当前批次损失值独立计算并实时写入优化器状态(如图1所示)。

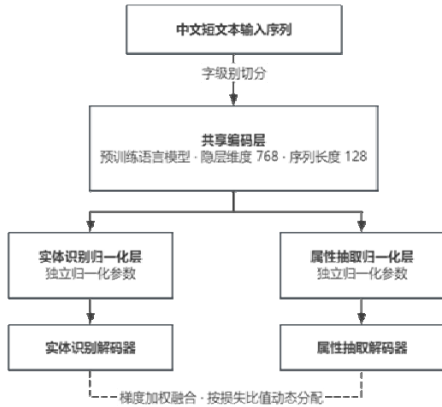


图1 基于预训练语言模型的跨任务共享编码层结构图

2.2 基于边界感知的短文本实体识别序列标注设计

实体识别模块采用序列标注范式, 在标准BIO标注体系基础上扩展为BIOES体系, 将单字实体、实体起始、实体内部、实体结尾及非实体五类标签精细区分, 标签类别数量扩展至实体类型数量的4倍加1。针对中文短文本上下文稀疏导致的边界模糊问题, 在编码层输出的基础上引入跨度感知注意力机制, 对每个候选实体跨度计算长度为1至6的边界置信度分布, 将跨度级别的边界特征与token级别的标注特征进行拼接融合, 边界特征维度设定为256。解码阶段采用条件随机场对标注序列施加全局转移约束, 转移矩阵参数在训练过程中与编码层参数联合更新, 通过维特比算法在长度不超过128的短文本序列上完成最优标注路径的搜索, 实体识别模块对每个识别实体跨度的token表示进行边界加权聚合后存储为固定维度向量, 作为后续跨模块传递的结构化输入单元。

2.3 基于实体驱动的短文本属性-值对解码设计

属性抽取模块以实体识别模块输出的实体跨度表示为驱动锚点, 将属性抽取任务建模为以实体为查询的跨度抽取问题, 针对每个识别实体在原始序列上定位与之对应的属性值文本跨度^[4]。实体跨度表示通过对实体内部token向量求加权均值获得, 权重由实体内部各token与实体中心token的注意力得分决定, 实体表示维度与编码层保持一致为768。解码器采用双指针网络结构, 以实体跨度表示为查询向量对序列中每个位置分别计算属性值起始概率与终止概率, 两组概率分布通过128维的属性类型嵌入向量进行条件调制, 属性类型嵌入向量在训练过

程中与解码器参数联合优化, 针对短文本高信息密度场景下多属性并存的特点, 各属性类型嵌入向量之间引入正交约束损失项以强化类型间的表示差异。解码器支持单实体多属性的并行抽取, 通过属性类型掩码矩阵对无效属性类型的解码路径进行屏蔽, 将每次前向传播的解码路径数量控制在预定义属性类型数量范围内。

2.4 实体识别对属性抽取的单向跨模块语义传递机制

跨模块语义传递机制以实体识别模块的中间层语义表示为传递载体, 将实体边界感知过程中积累的上下文语义信号以单向方式注入属性抽取模块的解码初始状态, 在两任务之间建立区别于流水线范式的显式语义通路。传递载体通过提取条件随机场层之前的隐层输出, 对每个实体跨度内的token表示依据BIOES标注概率分布中起始标签与终止标签的置信度进行边界加权聚合, 聚合后的实体语义向量维度为512。语义注入采用门控融合机制, 以实体语义向量与属性抽取解码器初始隐层状态为输入, 门控融合过程如下:

$$h_0^{inject} = \sigma(W_g \cdot e_{span}) \odot e_{span} + (1 - \sigma(W_g \cdot e_{span})) \odot h_0^{attr} \quad (1)$$

其中 e_{span} 为实体语义向量, h_0^{attr} 为属性抽取解码器的初

始隐层状态, h_0^{inject} 为门控融合后注入属性抽取模块的增强隐

层状态, W_g 为可学习融合门参数, σ 为sigmoid激活函数, $\odot$

为逐元素乘积, 门控参数在联合训练过程中依据两任务的损失动态比值自适应调整。单向传递方向基于两任务的逻辑依存关系, 通过切断属性抽取至实体识别的反向梯度通路规避双向耦合带来的训练不稳定风险。

2.5 针对短文本样本稀疏与标注噪声的联合优化目标设计

联合优化目标由实体识别损失与属性抽取损失两部分构成, 两者通过动态权重加权求和形成统一的训练目标函数, 联合损失定义如下:

$$L = \lambda_1 \cdot L_{NER} + \lambda_2 \cdot L_{AE} \quad (2)$$

其中L为整体联合损失, L_{NER} 为实体识别损失, L_{AE} 为属

性抽取损失, λ_1 与 λ_2 为动态权重系数, 两者之和在每个训练步归一化处理后写入优化器调度器, 并依据两任务当前训练步的验证集损失比值实时更新^[5]。实体识别损失采用条件随机场负对数似然函数, 对边界标签赋予1.5倍的损失权重, 加权系数依据训练集中边界标签与非边界标签的比例倒数动态校准, 属性抽取损失采用二元交叉熵函数分别计算起始指针与终止指针的预测误差, 对跨度长度超过8个token的预测结果施加长度惩罚项, 惩罚系数随跨度长度超出阈值的幅度线性递增。两项损失函

数中均引入标签平滑机制,平滑系数设定为0.1,将硬标签转化为以真实标签为峰值、以均匀分布为底的软概率分布。

3 实验验证与结果分析

3.1面向中文短文本场景的数据集、基线模型与评价指标

实验数据集选取面向中文短文本场景的公开语料,涵盖电商评论及社交媒体两类典型短文本领域,训练集、验证集及测试集按8:1:1比例划分,平均句长控制在25字以内,标注体系采用BIOES序列标注规范,属性-值对标注覆盖六类实体类型及对应属性类别。基线模型选取四类对比方法,分别为基于规则的流水线方法、仅含实体识别的单任务模型、仅含属性抽取的单任务模型以及已有联合抽取基线模型,所有基线模型均在相同数据划分条件下训练与测试。评价指标采用精确率与召回率及F1值三项标准指标,实体识别与属性抽取两个子任务分别独立计算,以宏平均F1值作为整体性能的综合衡量标准,同时报告严格匹配与部分匹配两种评价口径下的结果,以全面反映模型在短文本本边界模糊场景下的识别能力。

3.2联合模型与对比方法的性能分析

联合模型在实体识别与属性抽取两项子任务上均取得优于各基线方法的性能表现,尤其在短文本边界识别精度与属性-值对完整抽取率方面提升显著,各方法在测试集上的详细性能对比如表1所示。

表1 各方法在测试集上的性能对比(%)

方法	实体识别 P	实体识别 R	实体识别 F1	属性抽取 P	属性抽取 R	属性抽取 F1	F1提升
规则流水线方法	71.3	68.5	69.9	65.2	61.8	63.4	—
单任务实体识别模型	78.6	75.2	76.9	—	—	—	—
单任务属性抽取模型	—	—	—	72.4	69.3	70.8	—
联合抽取基线模型	80.1	77.4	78.7	74.6	71.2	72.8	—
已有最优联合方法	82.3	79.8	81.0	76.9	73.5	75.2	—
本联合模型	85.7	83.2	84.4	80.3	77.6	78.9	+3.4/+3.7

由表1可见,联合模型实体识别F1值较已有最优方法提升3.4个百分点,属性抽取F1值提升3.7个百分点,跨模块语义传递机制使属性抽取更充分利用实体识别结果,验证了联合建模框架相较流水线方法及单任务范式的优势。

3.3消融实验与关键模块有效性验证

消融实验逐一移除关键模块,对比与完整模型的性能差异。

移除共享编码层后,两任务F1值分别下降2.8与3.1个百分点; BIOES退化为BIO并去除跨度感知注意力后,实体识别F1值下降3.6个百分点,单字实体及紧邻实体误识别率上升;移除属性类型嵌入条件调制后,属性抽取F1值下降2.4个百分点;切断跨模块单向语义传递通路后,属性抽取F1值下降4.2个百分点,降幅最大;去除联合优化目标后,两任务F1值在低资源子集上下降尤为突出。各模块缺失均导致性能下滑,证实各组成部分均不可替代。

4 结语

面向中文短文本的实体识别与属性抽取联合模型,围绕短文本场景的核心挑战展开系统设计。跨任务共享编码、边界感知标注、实体驱动解码与单向跨模块语义传递四者协同运作,以显式的模块间信息通路取代流水线式独立建模范式,联合优化目标针对短文本噪声特性定向设计,强化了模型在低资源场景下的鲁棒性,消融实验证实各模块缺失均导致性能下滑,整体框架的有效性得到充分验证。未来可进一步探索跨域短文本场景下的模型迁移能力,以及在弱监督条件下扩展标注规模的可行路径。

【参考文献】

[1]王小龙,王暖臣,穆歌,等.基于数据增强的作战体系命名实体识别研究[J].舰船电子工程,2025,45(05):23-27+106.
[2]周烁,方正,蒋明鹏,等.基于大型语言模型的中文短文本实体链接方法[J].指挥信息系统与技术,2024,15(06):41-47.
[3]夏旭东,于荣欢.基于BERT+Bi-LSTM+CRF的航天领域命名实体识别研究[J].兵工自动化,2024,43(02):78-83+92.
[4]周贤琛,马扬,程光权,等.战场目标实体类型识别的鲁棒图神经网络方法[J].智能系统学报,2023,18(06):1156-1164.
[5]王秀慧,徐永波.软注意力掩码嵌入下中文命名实体识别算法[J].吉林大学学报(工学版),2026,56(01):231-238.

作者简介:

展辉辉(1988--),男,汉族,河南省新蔡县人,硕士研究生,讲师,研究方向: 电子信息技术。
叶佩雯(1997--),女,汉族,湖北省安陆市人,硕士研究生,助教,研究方向: 计算机科学与技术。
李默宇(1998--),男,汉族,黑龙江省绥化市人,硕士研究生,助教,研究方向: 电子信息技术。