doi:10.11887/j.cn.202404019 http://journal.nudt.edu.cn

卫星领域语料库构建与命名实体识别

徐 聪[1,2],石会鹏[3],陈志敏[1],张鑫宇[1,2],王 静[1],杨甲森[1*]
(1. 中国科学院国家空间科学中心 复杂航天系统电子信息技术重点实验室,北京 100190;
2. 中国科学院大学,北京 100049;3. 国家无线电监测中心检测中心,北京 100041)

摘 要:针对卫星领域命名实体语料匮乏、现有算法识别性能较低的问题,提出一种考虑模糊边界的卫星领域实体标注方法,构建包含8类常见卫星领域实体的语料库,与该领域现有语料库相比粒度更细、覆盖更广,并以此为基础提出迁移学习和多网络融合的卫星领域实体识别算法。该算法采用预训练双向编码器对语料语义平滑迁移获得子词级别特征,采用双向长短期记忆(bi-directional long-short term memory, BiLSTM)神经网络捕捉上下文信息确定边界,以条件随机场作为解码器实现标签预测。实验结果表明:相比于BiLSTM等传统模型具有更优的识别性能,算法在8种实体上的F_1值均在92%以上,微平均F_1值达到96.10%。

关键词:命名实体识别;迁移学习;神经网络;数据稀缺
中图分类号:V419; TP391.1 **文献标志码**:A **文章编号**:1001-2486(2024)04-175-09

论文拓展

Satellite domain corpus construction and named entity recognition

XU Cong[1,2], *SHI Huipeng*[3], *CHEN Zhimin*[1], *ZHANG Xinyu*[1,2], *WANG Jing*[1], *YANG Jiasen*[1*]
(1. Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. The State Radio_monitoring_center Testing Center, Beijing 100041, China)

Abstract: Aiming at the lack of named entity corpus in the satellite domain and the low recognition performance of existing algorithms, a satellite domain entity labeling method considering fuzzy boundaries was proposed, constructed a corpus containing 8 common satellite domain entities where the granularity was finer and the coverage was wider in comparison with the existing corpora in this field. Based on this, a transfer learning and multi-network fusion satellite domain entity recognition algorithm was proposed. Algorithm used pretrained bidirectional encoder representations for transformers to smoothly transfer the semantics of the corpus for subword-level features, a BiLSTM (bi-directional long-short term memory) network for capturing contextual information to determine boundaries, and label prediction was achieved using a conditional random field as a decoder. Experimental results show that, compared with traditional models such as BiLSTM, the proposed algorithm has better recognition performance where the F_1-score in 8 entities is all above 92% and the micro-average F_1-score reaches 96.10%.

Keywords: name entity recognition; transfer learning; neural networks; data scarcity

卫星情报具有敏感性、隐蔽性、稀缺性[1-2],互联网公开数据蕴含丰富的高价值卫星情报。卫星领域命名实体识别(name entity recognition, NER)是从互联网公开数据中自动识别在轨卫星名称、发射机构、政府组织、有效载荷、频率轨道等实体,为卫星情报知识图谱构建[3]、频率轨道资源全球态势评估[4]、战略意图窥探[5]等提供实体基础的一项工作。

现有命名实体识别方法主要有3类。基于规则和字典的方法[6]识别召回率较高但必须依托大量人工成本来制作规则和词典;隐马尔可夫模型(hidden Markov model, HMM)和条件随机场(conditional random field, CRF)[7]等基于特征工程的方法可以减少对规则和词典的依赖,但领域

收稿日期:2022-04-15
基金项目:中国科学院复杂航天系统电子信息技术重点实验室择优基金资助项目(Y42613A32S)
第一作者:徐聪(1997—),女,山东济宁人,博士研究生,E-mail:xucong19@mails.ucas.edu.cn
*通信作者:杨甲森(1979—),男,山东聊城人,研究员,博士,博士生导师,E-mail:jsy@nssc.ac.cn
引用格式:徐聪,石会鹏,陈志敏,等. 卫星领域语料库构建与命名实体识别[J]. 国防科技大学学报, 2024, 46(4): 175-183.
Citation: XU C, SHI H P, CHEN Z M, et al. Satellite domain corpus construction and named entity recognition[J]. Journal of National University of Defense Technology, 2024, 46(4): 175-183.

知识和人工设计特征模板受限于特定领域,迁移效果不佳;近年来,基于数据驱动的深度学习方法以其无须大量人工投入就能自动发现隐藏特征的优势,在通用领域[8]和生物医药[9]、金融风险[10]、法律文书[11]等特定领域被广泛应用,此类方法的基础是获得用于模型训练的大规模领域语料库。

在SpaceX[12]、OneWeb[13]等公司推出的低轨互联网星座驱动下,卫星领域实体识别成为新的研究方向。Jafari等[14]于2021年提出端到端SatelliteNER模型,与现有NER工具相比,该模型在卫星数据集上的实体识别效果最好,F_1值达到70%。SatelliteNER模型的缺点也比较明显:①实体类型有限、粒度过粗、标注未考虑卫星领域实体边界模糊等特征。其卫星数据集仅包含组织机构、火箭名称和卫星名称3类;卫星名称实体没有细化至编号,例如"Starlink-31"和"Starlink-32"两颗卫星被统一缩减为"Starlink";实体标注未考虑边界模糊、实体嵌套等卫星领域实体特点。②模型识别性能尚存较大改进空间。在测试集上的F_1值仅为70%,召回率比精度低约10%。

针对卫星领域命名实体语料库匮乏,现有命名实体识别算法性能较低问题,在分析卫星领域命名实体特点的基础上,制定模糊边界卫星领域实体标注规则,采用最大后向匹配方法实现数据集自动标注,构建卫星语料库SatCorpus,并以此为基础提出预训练双向编码器-双向长短期记忆神经-条件随机场(bidirectional encoder representations for transformers-bi-directional long short-term memory-conditional random field, BERT-BiLSTM-CRF)融合的卫星领域实体识别方法。实验结果表明,相比于BiLSTM、BiLSTM-CRF、卷积神经网络(convolutional neural network,CNN)-BiLSTM、双向门控循环单元(bi-gate recurrent unit, BiGRU)-CRF、BERT-BiGRU-CRF等传统方法,BERT-BiLSTM-CRF模型具有更优的实体识别性能。

1 卫星领域命名实体特点分析

与通用领域和规范化的医疗病例、司法文书等特定领域实体识别研究不同,卫星领域命名实体具有以下特点:

1)嵌套特征突出。例如,"U.S. Air Force"分开看"U.S."属于来源地实体,"Air Force"属于政府组织实体,而整体词组属于政府组织实体。

2)专业名词符号化。由于卫星领域的独特性和专业性,语料中存在大量的符号化表述,例如发射场"Launch Complex 39A"由文字+数字+编号结合,"Soyuz-2.1b"和"Soyuz 2.1b"均表示联盟号2.1b型火箭。

3)实体隐蔽性强。卫星与军事领域相关度较高,执行任务和服务对象具有特殊性,实体指向性不明显。例如,通用领域命名实体识别方法从"USA 148"中识别出美国(USA),实际上"USA 148"是军事卫星名称。

4)实体粗细粒度不统一。例如,针对同一发射地点,有语料为"Kennedy space center",还有语料会具体到发射场的某一发射台,如"Kennedy space center launch complex 39A"。

5)实体表达多样性。存在实体简化表述现象,例如,不少语料会将政府组织"defense advanced research projects agency"缩写为"DARPA"。

6)实体分布稀疏。以卫星发射语料为例,一篇新闻的大部分单词和词组属于无效实体,卫星领域实体在整篇文章中分布稀疏。

以上6个特点除了"实体分布稀疏"属于卫星领域命名实体的统计特征外,其余5个特点导致卫星实体边界模糊、难以确定。在卫星领域,不仅命名实体复杂度高,还面临语料匮乏的难题。现有研究缺乏面向卫星领域的实体分类标准和公开数据集,阻碍深度学习方法的应用。

2 卫星领域语料库构建

针对卫星领域命名实体的特点,本节提出考虑模糊边界的实体标注规则和自动标注方法,基于航天新闻构建针对卫星领域8个类别共计20 110条句子的语料库SatCorpus。具体方法说明如下。

2.1 卫星领域命名实体与标注类别

卫星的生命周期大致可分为研制、发射、在轨、离轨四个阶段。论文主要聚焦于研制和发射阶段的卫星领域实体。

除卫星名称和通用的时间、来源地(地名)外,对运载火箭、航天机构、制造公司、政府组织、发射场所5类实体给出如下定义:将把卫星送入预定轨道的航天运载工具标注为运载火箭实体;将负责航空航天和深空探测等太空任务的研究院所和开发组织标注为航天机构实体;将生产与卫星相关的航天器制造商、推进器制造商、卫星发射器和航天原件制造商等企业标注为制造公司实体;将与卫星服务有关的作战部队、政府机关、行政单位等标注为政府组织实体;将与发射火箭相关的建筑、场地或设施标注为发射场所实体。

采用通用、高效的开始 - 内部 - 外部(begin, inside, outside,BIO)标注方法[15]对实体进行单词级别的序列标注,“B”表示实体的开始,“I”表示实体的内部,“O”表示非实体部分。实体示例与详细标注方式见表 1,“teleport”“antenna”等领域外实体被标注为“O”。

表 1　卫星领域命名实体与标注类别

Tab. 1　Satellite domain named entity label category

实体类别	实体举例	子词分割	标注示例
卫星名称	Starlink 31, CAPELLA 1	Starlink 31	B - SAT I - SAT
时间	May 18, Dec 21	May 18	B - TIME I - TIME
来源地	USA, China	USA	B - SOU
运载火箭	Falcon 9, Ares 5	Falcon 9	B - ROC I - ROC
航天机构	NASA, ESA	NASA	B - AGEN
制造公司	SpaceX, Lockheed Martin	SpaceX	B - COM
政府组织	Air Force, DARPA	Air Force	B - ORG I - ORG
发射场所	Baikonur Cosmodrome	Baikonur Cosmodrome	B - SITE I - SITE

2.2　考虑模糊边界的卫星领域实体标注规则

不同实体标注规则会引发实体识别效果的差异[15]。对于卫星名称、时间、来源地、运载火箭、航天机构、制造公司、政府组织、发射场所 8 类命名实体,时间实体规则性强,来源地实体表示的全球国家及地区名称有限,制造公司和航天机构实体具有独立性和唯一性,因此模糊边界主要存在于卫星名称、运载火箭、政府组织、发射场所实体中。针对以上卫星领域实体边界模糊、嵌套严重、隐蔽性强等特征,结合领域专家的知识,给出考虑模糊边界的卫星领域实体标注规则。

规则 1:短横线或空格键、数字与卫星名称相连,统一标注为卫星名称。例如“Starlink - 61”。

规则 2:短横线或空格键、数字与运载火箭相连,统一标注为运载火箭。例如“Long March 5”。

规则 3:国家与政府组织相连,将相连的整体标注为政府组织。例如“U. S. Air Force”。

规则 4:发射台与发射场所相连,如果发射台从属于发射场所,则认为发射台是发射场所实体的细粒度实体,标注为发射场所。例如“Launch Complex 39A at the Kennedy Space Center”中的“Launch Complex 39A”。

规则 5:国家与数字相连一般为军事卫星,标注为卫星名称。例如“USA 134”。

2.3　考虑模糊边界的卫星领域实体标注方法

基于考虑模糊边界的卫星领域实体标注规则构建各实体领域词典。若同一实体存在于多个词典中,模糊边界在实体的长度上表现最为直接,因此采用最大后向匹配方法将实体长度最大的领域词典对应的实体类型作为该实体标签,实现数据集自动标注。

定义(最大后向匹配法)　设 $w_1,w_2,\cdots,w_n$ 是待标注的单词,$D_1,D_2,\cdots,D_k$ 是 k 个实体类别 Tag_1, $Tag_2,\cdots,Tag_k$ 对应的领域词典,以单词 w_i 起始在 $D_1,D_2,\cdots,D_k$ 中匹配实体长度最大的领域词典所对应的实体类型标注为以 w_i 开始的实体类型,记

$$Tag_i = \arg\max_{1\leqslant j\leqslant k}(l_j) \tag{1}$$

式中,l_j 表示在词典 D_j 中匹配的实体长度。基于最大后向匹配法的数据集自动标注算法流程,见表 2。

表 2　基于最大后向匹配法的数据集自动标注算法

Tab. 2　Automatic labeling algorithm of datasets based on maximum backward matching method

输入:待标注文本单词序列 $S=(w_1,w_2,\cdots,w_n)$,
领域词典集合 $D=\{D_1,D_2,\cdots,D_k\}$,
实体类别集合 $Tag=\{Tag_1,Tag_2,\cdots,Tag_k\}$
输出:待标注文本标签集合 $T=\{T_1,T_2,\cdots,T_n\}$

```
T←Ø                          % 初始化集合为空
for i = 1 to n do
    tmp←Ø                    % 初始化集合为空
    for j = 1 to k do        % 最大后向匹配法
        l_j←以单词 w_i 为开始在词典 D_j 中匹配的实体长度
        tmp←tmp∪{l_j}
        j←j+1
    end
    p = arg max(tmp)         % p ∈ {1,2,…,k}
    T_i = Tag_p
    T←T∪{T_i}
    i←i+l_p
end
```

2.4 SatCorpus 语料库构建

卫星领域语料库 SatCorpus 的语料来源于 SpaceNews 卫星发射新闻网(https://spacenews.com/)。语料库构建分为新闻采集、数据清洗、自动标注、人工校对 4 个步骤,流程如图 1 所示。

1)新闻采集:选取 SpaceNews“发射”栏目下 2013 年 3 月 20 日至 2021 年 11 月 18 日间的 1 129篇新闻作为原始语料数据集。

2)数据清洗:删除图片、广告、网址等不包含卫星信息的文字,以“,”“.”“?”“!”等标点符号作为标志对清洗后的数据进行句子级别的划分。

3)自动标注:确定面向卫星领域命名实体的划分类别,考虑实体的模糊边界和简化表达,采用最大后向匹配算法对未经标注的原始语料库进行标注,删除不包含实体的句子。

4)人工校对:人工核实并更正标注结果。

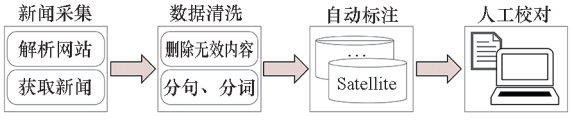

图 1 SatCorpus 卫星语料库自动标注流程

Fig. 1 Satellite corpus SatCorpus automatically annotation process

论文从 31 162 条原始语料数据集中,构建包含 20 110 条句子、8 个实体类别的语料库 SatCorpus,每种类别的实体数量见表 3。

表 3 SatCorpus 语料库实体数量

Tab. 3 Entity number in the SatCorpus corpus

实体类别	实体符号	实体数量
卫星名称	SAT	15 397
时间	TIME	6 217
来源地	SOU	4 795
发射场所	SITE	1 656
运载火箭	ROC	5 522
航天机构	AGEN	1 919
制造公司	COM	6 004
政府组织	ORG	3 443

相比于 Jafari 等[14]构造的只考虑 3 个实体类别的语料库,语料库 SatCorpus 实体覆盖度更广,粒度更细致,且考虑模糊边界和实体表达多样性,更符合卫星领域实体特点。

3 卫星领域实体识别模型

卫星领域实体分布稀疏,所构建语料库 SatCorpus 规模略小,论文采用适用于小数据集的迁移学习方法、多网络分层模型(BERT-BiLSTM-CRF)进行卫星领域命名实体识别。该模型由基于迁移学习的 BERT 输入表示层、基于 BiLSTM 的文本编码层、基于 CRF 的标签解码层三部分组成。图 2 展示了实体识别模型的整体结构,矩形结构表示分词环节,正方形结构表示向量,圆角矩形结构表示子算法,圆形结构表示字符串,其中加粗圆形结构对应输入输出。

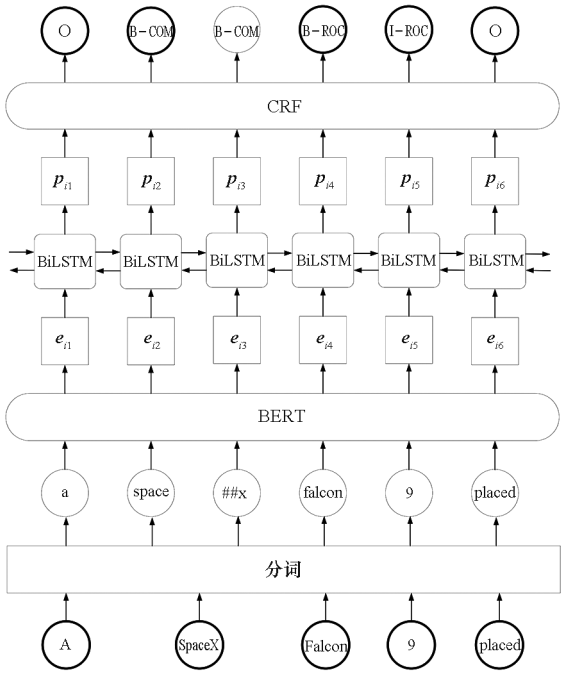

图 2 卫星领域命名实体识别模型结构

Fig. 2 Named entity recognition model structure in satellite domain

记 $S=(s_1,s_2,\cdots,s_n)$ 表示包含 n 条句子的卫星语料库 SatCorpus,对句子的每个单词进行子词分割,$s_i=(w_{i1},w_{i2},\cdots,w_{im})$ 表示由 m 个子词构成的第 i 条句子,BERT 模型将每个子词 w_{ij} 表示为单词特征、字符特征以及位置特征之和的特征向量 $\boldsymbol{c}_{ij}$;以句子为单位,句子 s_i 的特征向量序列 $\boldsymbol{c}_i=[\boldsymbol{c}_{i1},\boldsymbol{c}_{i2},\cdots,\boldsymbol{c}_{im}]$ 经过 BERT 中的 Transformer 层将转化为 $\boldsymbol{E}_i=[\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{im}]$;经过 BiLSTM 模型正向和反向传播提取上下文特征后形成发射矩阵 $\boldsymbol{P}_i=[\boldsymbol{p}_{i1},\boldsymbol{p}_{i2},\cdots,\boldsymbol{p}_{im}]$;CRF 模型学习标签间的依赖关系,并结合发射矩阵 $\boldsymbol{P}_i$ 预测每个子词对应的标签,将实体所在的第一个子词的标签作为该实体最终标签。

3.1 基于 BERT 的输入表示层

卫星领域命名实体语料匮乏,通常考虑使用

迁移学习将已有足够训练数据的其他领域知识迁移到标记数据匮乏的目标领域上进行领域适应来改进性能。参数共享是迁移学习的方法之一[16],它假设源任务和目标任务共享一些参数或模型超参数的先验分布,经过预训练的 BERT 模型即采用参数共享提高模型的适用度[17],其预训练语料来自 BooksCorpus(8 亿个单词)和英文维基百科(25 亿个单词)[18]。BERT 模型在卫星领域的迁移学习路径如图 3 所示。

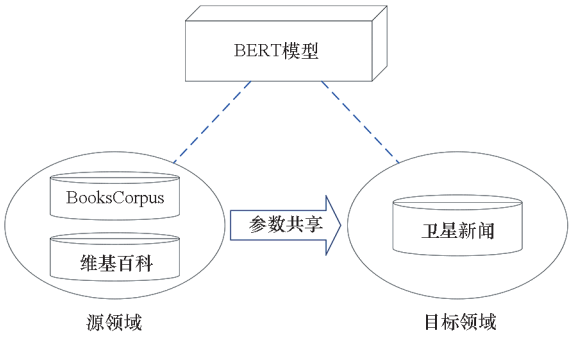

图 3　BERT 模型在卫星领域的迁移学习路径

Fig. 3　Transfer learning path of BERT model in satellite domain

基于迁移学习的 BERT 模型生成子词特征向量过程如下:对语料库第 i 条句子 s_i 的每个单词按照预训练模型的词汇表进行子词分割,例如单词“SpaceX”被划分为子词“space”和子词“##x”。固定特征向量长度为 l,含有 m 个子词的句子 $s_i=(w_{i1},w_{i2},\cdots,w_{im})$ 以子词为单位计算 3 个特征:词特征 $\boldsymbol{C}_1=[\boldsymbol{c}_{11},\boldsymbol{c}_{12},\cdots,\boldsymbol{c}_{1m}]\in\mathbf{R}^{l\times m}$、句子特征 $\boldsymbol{C}_2=[\boldsymbol{c}_{21},\boldsymbol{c}_{22},\cdots,\boldsymbol{c}_{2m}]\in\mathbf{R}^{l\times m}$、位置特征 $\boldsymbol{C}_3=[\boldsymbol{c}_{31},\boldsymbol{c}_{32},\cdots,\boldsymbol{c}_{3m}]\in\mathbf{R}^{l\times m}$。基于 BERT 的输入表示层的输入为三者数值之和 $\boldsymbol{C}_0=\boldsymbol{C}_1+\boldsymbol{C}_2+\boldsymbol{C}_3$,$\boldsymbol{C}_0=[\boldsymbol{c}_{01},\boldsymbol{c}_{02},\cdots,\boldsymbol{c}_{0m}]\in\mathbf{R}^{l\times m}$ 经过多层 Transformer[2] 输出最终的特征向量矩阵 $\boldsymbol{E}_i=[\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{im}]$,$\boldsymbol{E}_i\in\mathbf{R}^{l\times m}$。

3.2　基于 BiLSTM 的文本编码层

卫星领域命名实体识别同自然语言处理其他领域一样与上下文语境关联密切,具有长期依赖特点。长 - 短期记忆网络(long short-term memory,LSTM)凭借独特的记忆单元结构捕捉序列的长程依赖性。

对于第 i 条句子 $s_i=(w_{i1},w_{i2},\cdots,w_{im})$,BERT 的输出 $\boldsymbol{E}_i=[\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\cdots,\boldsymbol{e}_{im}]$,$\boldsymbol{E}_i\in\mathbf{R}^{l\times m}$ 作为 BiLSTM 层的输入。整个句子经过前向 LSTM 和后向 LSTM 计算每个单词的前向特征 $\boldsymbol{F}_i=[\boldsymbol{f}_{i1},\boldsymbol{f}_{i2},\cdots,\boldsymbol{f}_{im}]$ 和后向特征 $\boldsymbol{B}_i=[\boldsymbol{b}_{i1},\boldsymbol{b}_{i2},\cdots,\boldsymbol{b}_{im}]$,$\boldsymbol{F}_i,\boldsymbol{B}_i\in\mathbf{R}^{d\times m}$,二者拼接融合上下文特征,记为 $\boldsymbol{Q}_i=[\boldsymbol{F}_i;\boldsymbol{B}_i]\in\mathbf{R}^{2d\times m}$,最后经 softmax 层线性映射到标签空间 $\boldsymbol{P}_i\in\mathbf{R}^{k\times m}$,$k$ 为标签总数。

$$\boldsymbol{Q}_i=\begin{bmatrix}\boldsymbol{F}_i\\ \boldsymbol{B}_i\end{bmatrix}=\begin{bmatrix}\boldsymbol{f}_{i1} & \cdots & \boldsymbol{f}_{im}\\ \boldsymbol{b}_{i1} & \cdots & \boldsymbol{b}_{im}\end{bmatrix}=[\boldsymbol{q}_{i1}\quad \boldsymbol{q}_{i2}\quad \cdots\quad \boldsymbol{q}_{im}] \tag{2}$$

$$\boldsymbol{p}_{ij}=\mathrm{softmax}(\boldsymbol{W}\boldsymbol{q}_{ij}+\boldsymbol{b})\quad j=1,2,\cdots,m \tag{3}$$

式中,$\boldsymbol{p}_{ij}\in\mathbf{R}^{k\times 1}$ 表示第 i 条句子的第 j 个词分别对应 k 个标签的概率矩阵,$\boldsymbol{W}$ 表示权重矩阵,$\boldsymbol{b}$ 表示偏置向量,激活函数为 softmax 函数。由式(3)可知,softmax 层的输出是相互独立的,与相邻输出毫无影响,这与自然语言的连通性相违背,因此下一步引入 CRF 学习相邻标签的依赖关系。

3.3　基于 CRF 的文本解码层

卫星新闻中单词只与相邻的上下文有关,与其他内容无关,符合马尔可夫随机场的特征。条件随机场是给定一组输入随机变量对输出随机变量预测的条件概率模型,要求输出随机变量满足马尔可夫随机场。CRF 充分考虑相邻实体标签的依赖关系,基于这些约束从统计方法上提高预测标签序列的准确概率。

对于句子 $s=(w_1,w_2,\cdots,w_m)$,BiLSTM 层的输出 $\boldsymbol{P}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m]$ 作为 CRF 层的输入观测序列 $\boldsymbol{X}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m]$,设 CRF 层的输出标签序列 $\boldsymbol{y}=[y_1,y_2,\cdots,y_m]$,$y_i\in\{1,2,\cdots,k\}$ 表示 k 个标签,条件概率分布为线性链条随机场。

定义状态转移矩阵 $\boldsymbol{A}=\begin{bmatrix}a_{11} & \cdots & a_{1k}\\ \vdots & & \vdots\\ a_{k1} & \cdots & a_{kk}\end{bmatrix}$,$a_{ij}$ 表示标签 i 到标签 j 的转移概率,$\sum_{t=1}^{k}a_{it}=1$。定义特征函数:

$$s(\boldsymbol{X},\boldsymbol{y})=\sum_{i=1}^{m}a_{y_i,y_{i+1}}+\sum_{i=1}^{m}p_{y_i,i} \tag{4}$$

特征函数 $s(\boldsymbol{X},\boldsymbol{y})$ 第一项表示转移特征,第二项表示状态特征。

在观测序列 $\boldsymbol{X}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_m]$ 的条件下,输出标记序列 $\boldsymbol{y}=[y_1,y_2,\cdots,y_m]$ 的条件概率具有如下形式:

$$P(\boldsymbol{y}\mid\boldsymbol{X})=\frac{\exp(s(\boldsymbol{X},\boldsymbol{y}))}{\sum_{y}\exp(s(\boldsymbol{X},\boldsymbol{y}))} \tag{5}$$

训练时对条件概率 $P(\boldsymbol{y}\mid\boldsymbol{X})$ 取对数并最大化,最大值对应的标记序列 y^* 即为输出最优标签序列。

4 实验设计及结果分析

为了验证卫星领域实体识别模型的有效性，分别采用多模型对比和领域内对比的分析方法。多模型对比指的是在同一数据集下，与 BiLSTM、BiLSTM-CRF、BIGRU-CRF、CNN-BiLSTM-CRF 等常用实体识别模型对比分析；领域内对比指的是与卫星领域其他命名实体识别方法进行比较，包括数据集对比和模型对比。

4.1 评价方法

命名实体识别是信息抽取的研究内容之一，评价指标采用在消息理解系列会议（message understanding conference，MUC）中规定的衡量信息抽取系统性能的两个指标：精确率（P）和召回率（R）。为了综合评价抽取系统的性能，通常还计算二者的调和平均数——F_1 值[19]。

$$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R} \tag{6}$$

微平均（micro-average）F_1 值将各类实体依据实体数目先统计汇总再计算相应指标，有效缓解语料库 SatCorpus 中 8 类实体数据不均衡的现象。

4.2 实验设置

按照 6∶2∶2 的比例将语料库 SatCorpus 划分为训练集、验证集、测试集，采用准确率、召回率、F_1 值作为评价指标。实际实验中预训练 BERT 模型采用含有 12 个 Transformer 隐藏层、12 个多头注意力和 768 维隐藏层的 BERT-base 模型。由于 BERT 模型以子词为单位输出特征向量，在文本解码层取实体所含子词的第一个标签作为预测标签。实验基于 TensorFlow2.3.0 深度学习框架在单 NVIDIA GeForce RTX 2080 Ti 上训练模型，模型参数设置见表 4。

表 4 BERT-BiLSTM-CRF 模型实验参数

Tab.4 BERT-BiLSTM-CRF model experimental parameters

实验参数	取值
最大句子长度	50
批处理大小	64
训练次数	30
失活率	0.4
编码维度	768

4.3 实验效果分析

BERT-BiLSTM-CRF 模型对卫星领域不同实体的识别结果见表 5。其中对于“制造公司”和“时间”两类实体的识别效果达到 99% 以上。由于“时间”实体的规则性强，与特定词（如“year”“month”等）存在很强的关联度，模型对这类实体的识别精度高。模型对“卫星名称”实体的识别性能相对较差，召回率比精确率下降接近 4%，主要是由于该类实体结构较复杂，存在名称歧义性严重等问题，例如“First”卫星的名称易与常见英文单词“first”混淆，导致将其判别为非卫星实体而降低召回率。

表 5 基于 BERT-BiLSTM-CRF 的卫星领域实体识别结果

Tab.5 Entity recognition results in satellite domain based on BERT-BiLSTM-CRF

实体类别	精确率/%	召回率/%	F_1 值/%	测试集实体数量
来源地	96.16	96.45	96.31	987
卫星名称	94.19	90.67	92.39	3 021
制造公司	99.57	98.47	99.02	1 177
运载火箭	97.93	97.49	97.71	1 116
时间	99.52	99.44	99.48	1 261
发射场所	96.69	96.42	96.55	363
政府组织	98.24	96.82	97.53	692
航天机构	97.78	95.14	96.44	370
微平均值	96.91	95.29	96.10	8 987

多模型对比：为了进一步验证 BERT-BiLSTM-CRF 模型在卫星领域实体识别效果的优异性，在同一语料库 SatCorpus 上分别对 BiLSTM、BiLSTM-CRF、CNN-BiLSTM、BiGRU-CRF、BERT-BiGRU-CRF 模型进行实验。对上述模型按照输入表示层、文本编码层、文本解码层进行网络结构划分，前四个模型均没有特定的输入表示层，而是随机初始化子词向量，BiLSTM 模型没有明确地划分出文本编码层和标签解码层。各模型的微平均 F_1 值见表 6 和图 4。

表 6 命名实体识别模型效果对比

Tab.6 NER model effect comparison

输入表示	上下文编码器	标签解码器	F_1 值/%
	BiLSTM		93.17
	BiLSTM	CRF	94.36
	CNN	BiLSTM	94.52
	BiGRU	CRF	95.37
BERT	BiGRU	CRF	95.53
BERT	BiLSTM	CRF	96.10

由图 4 可知,BERT-BiLSTM-CRF 模型 F_1 值在 8 类实体的 6 类中均处于较高水平,对“卫星名称”实体识别精度超过 92%,在“来源地”实体中几乎接近于最优值,说明该模型在卫星领域的实体识别性能高于其他模型。不同模型结合不同实体的对比分析如下:

1)相比于其他模型,BiLSTM 缺少明确的输入表示层和文本解码层,不仅微平均 F_1 值在最低水平,在不同实体的识别精度上也基本处于最低位,对“发射场所”实体的识别精度低于 76%。即便只有单一结构,该模型在“时间”实体上的 F_1 值也高达 98%,这说明对于规则性强的实体,弱表达模型的识别能力不逊色于强表达模型。

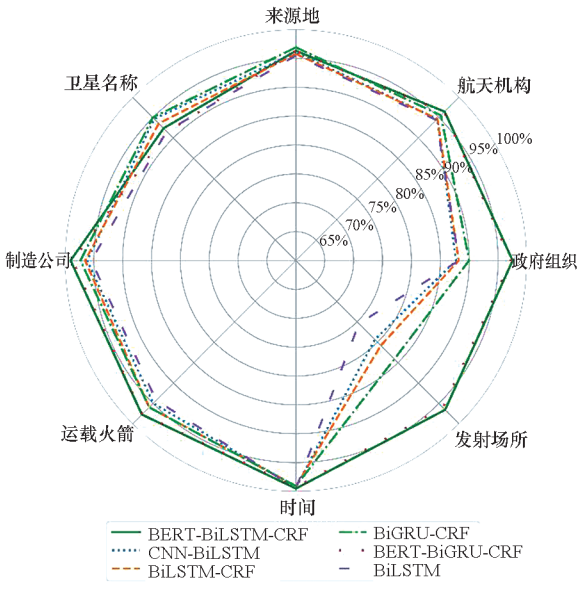

图 4　不同实体在不同模型上的 F_1 值

Fig. 4　F_1-score for different entities on different models

2)相比于 BiLSTM 模型,BiLSTM-CRF 模型的微平均 F_1 值提高 1.19%。由于 CRF 是基于单词特征和转移概率学习语料的特征,更易于抽取单词间的依赖关系,如“O”标签后的下一个实体只能以“B-”而不是“I-”开头。除了“时间”实体外,其他实体的 F_1 值均有明显提升,“发射场所”实体的 F_1 值提升幅度最大,达到 5.37%。

3)相比于 BiLSTM-CRF 模型,BERT-BiLSTM-CRF 模型的微平均 F_1 值提高 1.74%,具体表现为“发射场所”和“政府组织”两类实体的 F_1 值分别增加 15.71% 和 9.34%。这两类实体如 2.1 节所述嵌套严重,且表达方式较为灵活,如“发射场所”实体有文字 + 数字 + 编号(Launch Complex 39A)结合和纯文字描述(Kennedy Space Center)两种表达方式,“政府组织”实体存在全称(defense advanced research projects agency)与缩写(DARPA)两种不同的描述方法。

4)BiGRU-CRF 模型比 BiLSTM-CRF 模型的整体 F_1 值提高 1.06%,BERT-BiGRU-CRF 模型反而比 BERT-BiLSTM-CRF 模型降低了 0.57%。这是因为 BiGRU 简化了 BiLSTM 结构,在接入庞大的预训练模型后对基于 BERT 输入表示向量的上下文表征能力弱于 BiLSTM。

5)不同模型在不同实体上的 F_1 值在“发射场所”和“政府组织”两类实体上差异显著。对于上述两个实体类型,BERT-BiLSTM-CRF 模型略高于 BERT-BiGRU-CRF 模型,领先 BiGRU-CRF 模型 12.5% 和 7.54%,大幅领先 BiLSTM 模型 21.11% 和 9.29%。

领域内对比:与基于 spaCy 库的一种使用线性统计学习方法的卫星领域命名实体识别模型 SatelliteNER[14] 相比,本方法应用迁移学习方法解决标注数据匮乏的问题,能够识别更多种类实体,召回率和精度实现均衡,见表 7。

表 7　与现有卫星领域命名实体识别方法对比

Tab. 7　Comparison with the existing satellite domain named entity recognition methods

命名实体识别方法	数据集			实验效果		
	实体类别/个	数据集规模/句	是否考虑模糊边界	精度/%	召回率/%	F_1 值/%
SatelliteNER[14]	3	25 628	否	约 82	约 62	约 72
BERT-BiLSTM-CRF	8	20 110	是	96.88	95.29	96.10

在其他领域的命名实体识别研究上,BioBERT[20] 将预训练模型 BERT 应用于生物医学语料库;BERT-BiGRU-Attention-CRF [21] 在编码层与解码层之间添加注意力层增加对企业年报上下文本的表征能力;FMM-CAM[22] 针对法律文书实体名称长、关联度大的特点,建立匹配词分区定位词的边界。本文模型相比于上述模型有如下特点:①除使用预训练 BERT 模型构建输入表示层

外,还使用 BiLSTM 和 CRF 分别构建编码层和解码层,结构清晰,提升模型的可读性;②BERT 是一种基于 Transformer 的预训练模型,Transformer 的核心是多头注意力机制——以类似自相关和互相关的形式允许单词获得邻域单词的注意力分布从而储存长期记忆信息,鉴于 BERT 自带注意力机制,因此后续模型搭建中没有再过多添加注意力层,简化了模型结构;③通过构建模糊边界规则实现边界定位,识别效果更高,在三种领域方法中 F_1 值最高。

综上所述,基于预训练 BERT 的输入表示层弥补了在小数据卫星领域中特征不足的缺陷,BERT 的微调机制与后续 BiLSTM 和 CRF 层联动,提高对嵌套严重、边界模糊、表达方式多样的实体识别效果。

5 结论

针对开放域中卫星领域实体识别存在实体边界模糊、嵌套严重、实体种类丰富、规律性差、隐蔽性强等问题:

1)提出考虑模糊边界的卫星领域实体标注规则,结合领域专家知识构建卫星领域实体分类策略,实现数据集自动标注,建立了基于航天新闻的卫星领域语料库 SatCorpus,与该领域现有语料库相比,粒度更细,覆盖更广。

2)提出基于 BERT-BiLSTM-CRF 的多神经网络融合的卫星领域实体识别模型,通过在语料库 SatCorpus 上的多模型对比及领域内对比,结果表明:提出的 BERT-BiLSTM-CRF 模型优于较成熟的5种模型,微平均 F_1 值高达 96.10%,在每类实体上的 F_1 值均高于 92%,尤其是在“发射场所”和“政府组织”两类实体上,BERT-BiLSTM-CRF 模型领先 BiGRU-CRF 模型 12.5% 和 7.54%;模型优于卫星领域其他命名实体识别方法,取得高精度的同时提高召回率,实现精度和召回率的均衡。

下一步工作计划对语料库 SatCorpus 进一步扩充,探索无监督或半监督方法,在保证实体识别效果的基础上对模型进行简化,并对卫星领域的关系抽取方法开展研究。

参考文献(References)

[1] 刘全, 葛新, 李健十, 等. 非静止轨道宽带通信星座频率轨道资源全球态势综述(上)[J]. 卫星与网络, 2020(增刊1): 66 – 69.
LIU Q, GE X, LI J S, et al. Global situation of non-geostationary orbital broadband communication constellation frequency orbital resources (part 1)[J]. Satellite & Network, 2020(Suppl 1): 66 – 69. (in Chinese)

[2] 刘全, 葛新, 李健十, 等. 非静止轨道宽带通信星座频率轨道资源全球态势综述(下)[J]. 卫星与网络, 2020(3): 60 – 69.
LIU Q, GE X, LI J S, et al. Global situation of non-geostationary orbital broadband communication constellation frequency orbital resources (part 2)[J]. Satellite & Network, 2020(3): 60 – 69. (in Chinese)

[3] 刘舆, 曾德贤, 胡远方, 等. 基于知识图谱的卫星情报分析方法研究[J]. 情报探索, 2021(11): 1 – 7.
LIU Y, ZENG D X, HU Y F, et al. Research on satellite intelligence analysis method based on knowledge graph[J]. Information Research, 2021(11): 1 – 7. (in Chinese)

[4] 韩朝晖, 田伟, 李烨, 等. 用数据为卫星频率轨道资源管理服务[J]. 中国无线电, 2017(2): 28 – 29.
HAN C H, TIAN W, LI Y, et al. Use data to serve the management of satellite frequency orbit resources[J]. China Radio, 2017(2): 28 – 29. (in Chinese)

[5] 刘垚圻, 李红光, 周一青, 等. 数字孪生卫星互联网: 架构与关键技术[J]. 天地一体化信息网络, 2022, 3(1): 62 – 71.
LIU Y Q, LI H G, ZHOU Y Q, et al. Digital twin satellite internet: architecture and key technologies[J]. Space-Integrated-Ground Information Networks, 2022, 3(1): 62 – 71. (in Chinese)

[6] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the web: an experimental study[J]. Artificial Intelligence, 2005, 165(1): 91 – 134.

[7] 钟志农, 刘方驰, 吴烨, 等. 主动学习与自学习的中文命名实体识别[J]. 国防科技大学学报, 2014, 36(4): 82 – 88.
ZHONG Z N, LIU F C, WU Y, et al. Chinese named entity recognition combined active learning with self-training[J]. Journal of National University of Defense Technology, 2014, 36(4): 82 – 88. (in Chinese)

[8] BAEVSKI A, EDUNOV S, LIU Y H, et al. Cloze-driven pretraining of self-attention networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 5360 – 5369.

[9] ZHANG S D, ELHADAD N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts[J]. Journal of Biomedical Informatics, 2013, 46(6): 1088 – 1098.

[10] ALVARADO J, VERSPOOR K M, BALDWIN T. Domain adaption of named entity recognition to support credit risk assessment[C]//Proceedings of the Australasian Language Technology Association Workshop, 2015: 84 – 90.

[11] DOZIER C, KONDADADI R, LIGHT M, et al. Named entity recognition and resolution in legal text[M]//FRANCESCONI E, MONTEMAGNI S, PETERS W. Semantic processing of legal texts. Berlin: Springer-Verlag, 2010: 27 – 43.

[12] MCDOWELL J C. The low earth orbit satellite population and impacts of the SpaceX Starlink constellation [J]. The Astrophysical Journal Letters, 2020, 892(2): L36.

[13] RADTKE J, KEBSCHULL C, STOLL E. Interactions of the space debris environment with mega constellations: using the example of the OneWeb constellation[J]. Acta Astronautica, 2017, 131: 55 - 68.

[14] JAFARI O, NAGARKAR P, THATTE B, et al. SatelliteNER: an effective named entity recognition model for the satellite domain [C]//Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2020: 100 - 107.

[15] 宗成庆, 夏睿, 张家俊. 文本数据挖掘[M]. 北京: 清华大学出版社, 2019.
ZONG C Q, XIA R, ZHANG J J. Text data mining[M]. Beijing: Tsinghua University Press, 2019. (in Chinese)

[16] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345 - 1359.

[17] HAN X, ZHANG Z Y, DING N, et al. Pre-trained models: past, present and future[J]. AI Open, 2021, 2: 225 - 250.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2020 - 09 - 12]. http://arxiv.org/abs/1810.04805v2.

[19] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003(10): 1 - 5, 66.
LI B L, CHEN Y Z, YU S W. Research on information extraction: a survey [J]. Computer Engineering and Applications, 2003(10): 1 - 5, 66. (in Chinese)

[20] SONG B S, LI F, LIU Y S, et al. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison[J]. Briefings in Bioinformatics, 2021, 22(6): bbab282.

[21] 张靖宜, 贺光辉, 代洲, 等. 融入 BERT 的企业年报命名实体识别方法[J]. 上海交通大学学报, 2021, 55(2): 117 - 123.
ZHANG J Y, HE G H, DAI Z, et al. Namedentity recognition of enterprise annual report integrated with BERT[J]. Journal of Shanghai Jiao Tong University, 2021, 55(2): 117 - 123. (in Chinese)

[22] 郭力华, 李旸, 王素格, 等. 基于匹配策略和社区注意力机制的法律文书命名实体识别[J]. 中文信息学报, 2022, 36(2): 85 - 92.
GUO L H, LI Y, WANG S G, et al. Name entity recognition in legal instruments based on matching strategy and community attention mechanism [J]. Journal of Chinese Information Processing, 2022, 36(2): 85 - 92. (in Chinese)