文章编号：1006-2467(2021)02-0117-07　　DOI：10.16183/j.cnki.jsjtu.2020.009

融入 BERT 的企业年报命名实体识别方法

张靖宜[1]，　贺光辉[1]，　代　洲[2]，　刘亚东[1]
（1. 上海交通大学 电子信息与电气工程学院，上海 200240；2. 南方电网物资有限公司，广州 510641）

摘　要：自动提取企业年报关键数据是企业评价工作自动化的重要手段.针对企业年报领域关键实体结构复杂、与上下文语义关联强、规模较小的特点，提出基于转换器的双向编码器表示-双向门控循环单元-注意力机制-条件随机场(BERT-BiGRU-Attention-CRF)模型.在 BiGRU-CRF 模型的基础上，首先引入 BERT 预训练语言模型，以增强词向量模型的泛化能力，捕捉长距离的上下文信息；然后引入注意力机制，以充分挖掘文本的全局和局部特征.在自行构建的企业年报语料库内进行实验，将该模型与多组传统模型进行对比.结果表明：该模型的 F_1 值(精确率和召回率的调和平均数)为 93.69%，对企业年报命名实体识别性能优于其他传统模型，有望成为企业评价工作自动化的有效方法.

关键词：命名实体识别；企业年报；BERT；注意力机制；双向门控循环单元
中图分类号：TP 391.1　　**文献标志码**：A

Named Entity Recognition of Enterprise Annual Report Integrated with BERT

ZHANG Jingyi[1]，　*HE Guanghui*[1]，　*DAI Zhou*[2]，　*LIU Yadong*[1]
(1. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 2. China Southern Power Grid Materials Co., Ltd., Guangzhou 510641, China)

Abstract: Automatically extracting key data from annual reports is an important means of business assessments. Aimed at the characteristics of complex entities, strong contextual semantics, and small scale of key entities in the field of corporate annual reports, a BERT-BiGRU-Attention-CRF model was proposed to automatically identify and extract entities in the annual reports of enterprises. Based on the BiGRU-CRF model, the BERT pre-trained language model was used to enhance the generalization ability of the word vector model to capture long-range contextual information. Furthermore, the attention mechanism was used to fully mine the global and local features of the text. The experiment was performed on a self-constructed corporate annual report corpus, and the model was compared with multiple sets of models. The results show that the value of F_1 (harmonic mean of precision and recall) of the BERT-BiGRU-Attention-CRF model is 93.69%. The model has a better performance than other traditional models in annual reports, and is expected to provide an automatic means for enterprise assessments.

Key words: named entity recognition; enterprise annual report; BERT; attention mechanism; BiGRU

收稿日期：2020-01-08
作者简介：张靖宜(1996-)，女，河南省南阳市人，硕士生，主要从事自然语言处理的研究.
通信作者：贺光辉，男，副研究员，电话(Tel.)：021-34204546-1054；E-mail：guanghui.he@sjtu.edu.cn.

命名实体识别作为自然语言处理领域中的一项重要技术，与关系抽取[1]、事件抽取[2]、问答系统等其他自然语言处理任务的基础相关. 其主要负责准确、自动识别指定语料中实体(专有名词或有价值的短语)的边界并划分实体类别. 对企业年报进行命名实体识别可获得企业的基本信息和财务数据，为企业评价系统提供数据支撑，有助于企业掌握行业发展现状和趋势、规划发展方向、评估合作伙伴等. 因此，准确识别企业年报中的命名实体是建立企业评价系统的重要途径.

目前，命名实体识别方法包括：基于规则和字典、统计机器学习和深度学习. 其中，基于规则和字典的方法需要手动建立知识库和字典，耗时长且移植性差. 基于统计机器学习的方法应用较广泛，如隐马尔科夫模型(HMM)、条件随机场(CRF)、最大熵模型(ME)等，但以上方法需要人工设定特征模板，对语料库的依赖性较大且对特征选取要求较高. 与基于统计机器学习的方法相比，基于深度学习的方法能自动获取语料特征，命名实体识别的性能更好. 由于命名实体的标签之间的依赖关系较强，所以 Huang 等[3]将双向长短时记忆网络(BiLSTM)和 CRF 结合，所得模型能够利用过去和将来的信息更好地挖掘上下文关系. Chiu 等[4-5]将 BiLSTM 和卷积神经网络(CNN)结合，所得模型能够更好地利用前、后缀的字符级特征，减少人工构造特征. Cho 等[6]提出了门控循环单元(GRU)，其比长短时记忆网络(LSTM)少一个门，结构更简单，训练速度更快. 王洁等[7]将字向量作为输入，利用 BiGRU-CRF 模型提取会议名称的语料特征，发现与 LSTM 相比，GRU 的训练时间减少了 15%. Bharadwaj 等[8]在 BiLSTM-CRF 模型的基础上融入注意力机制，使模型更关注于对当前输出贡献大的字符. Cao 等[9]提出利用对抗迁移学习框架进行命名实体识别，通过提取不同任务中的共享词边界信息并利用自注意力机制，学习句子中任意 2 个字符之间的依赖关系. Vaswani 等[10]提出了利用自注意力机制快速并行的一种包含编码器和解码器的转换器(Transformer)模型. Devlin 等[11]提出了能够更好获取字符、词语和句子级别关系特征的基于转换器的双向编码器表示(BERT)预训练语言模型.

对企业年报进行识别的难点主要如下：① 专业财务术语、企业名称实体繁多. 其中，如净利润、营业收入等财务术语的专业性较强；企业名称包括以“有限公司”“集团”等为尾的全称和仅包含企业名称关键信息的简称；② 数值信息多且数字实体的识别难度大，如“公司 2011 年末总资产和归属于上市公司股东的所有者权益分别为 170 704.67 万元和 156 542.08 万元”，需要正确识别出财务术语对应的数值信息及其单位；③ 财务数值相对于上年变化趋势的描述方式多变，如下降 10%、同上年持平等；④ 企业年报语料库规模较小，仅为 1998 年人民日报语料库的 19.28%.

针对以上问题，提出 BERT-BiGRU-Attention-CRF 融合模型. 在基础模型 BiGRU-CRF 上引入 BERT 预训练语言模型，并在大型语料库上进行预训练学习语义特征，补足企业年报语料库的特征，克服语料库规模小的问题. 同时，BERT 利用 Transformer 模型提升自身模型的抽取能力，能够更好地明确实体边界. 此外，在 BiGRU-CRF 模型中引入注意力机制，便于理解句子结构，从而充分挖掘上下文的语义信息，进一步提升实体的识别性能.

1 企业年报数据集的构建

目前，关于企业年报的命名实体识别方法的研究较少，且缺乏实验测试所需的典型数据集，因此本文利用网络爬虫技术抓取企业官方年报，自行构建和标注该领域的数据集. 具体构建步骤如下：

(1) 数据预处理. 利用正则表达式从每份年报中自动提取出“企业经营概况”标题下的语段.

(2) 实体类别确立. 构建企业评价系统需要从企业年报中获取企业的基本信息和经营状况. 其中，基本信息包括年份和企业名称共两类实体；通过阅读企业年报和利用词频-逆文档频次(TF-IDF)算法[12]提取关键词的方式，选取与“利润”和“收入”相关的财务指标概括企业的经营状况. 自行标注的实体共 7 大类，如表 1 所示.

表 1 企业年报实体

Tab. 1 Entities of enterprise annual report

实体类别	实体示例	实体符号
年份	2005 年、2016 年	YEAR
企业名称	平安银行股份有限公司	COM
财务术语	净利润、营业收入、主营业务收入	FIN
利润类数值	12 626.23 万元	PNUM
收入类数值	138.64 亿元	SNUM
利润类数值同上年变化趋势	增长 13.27%、扭亏为盈	PTREND
收入类数值同上年变化趋势	下降 23.12%、创下新高	STREND

（3）标注体系. 实验采用的标注体系为 BIO. 其中，B 代表实体的起始位置，I 代表实体中除起始位置的其他部分，O 代表非实体部分. 需要预测的实体共 15 小类，标注示例如表 2 所示.

表 2　标注示例

Tab. 2　A example of labeling

语料字符	标注符号	语料字符	标注符号
实	O	入	I-FIN
现	O	1	B-SNUM
营	B-FIN	7	I-SNUM
业	I-FIN	万	I-SNUM
收	I-FIN	元	I-SNUM

2　BERT-BiGRU-Attention-CRF 模型

本模型由 BERT 预训练语言模型、BiGRU 网络、注意力机制和 CRF 层构成. 首先，把输入字符的字向量、文本向量和位置向量之和作为 BERT 的输入. 利用 BERT 获取上下文语义信息，把融合语义后的输出向量输入到 BiGRU 网络进行编码，前向 GRU 网络学习未来特征，反向 GRU 网络学习历史特征. 然后，将挖掘得到的全局特征，即 t 时刻的隐藏状态（$\boldsymbol{h}_t$）作为输出，并利用注意力机制补足局部特征，预测出输入文本序列与标签之间的关系. 最后，利用 CRF 进行解码预测标签之间的合理性关系，输出最佳标签序列，模型结构如图 1 所示.

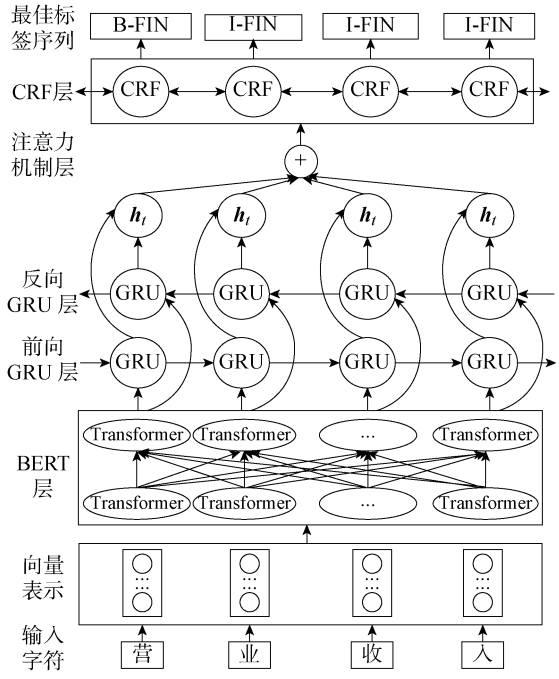

图 1　BERT-BiGRU-Attention-CRF 模型结构

Fig. 1　Structure of BERT-BiGRU-Attention-CRF model

2.1　BERT 预训练语言模型

BERT 预训练语言模型将深度学习的思想融入语言模型中，可将词表征为向量形式，从而获取词语之间的相似度[13]. 在双向 Transformer 编码器（见图 2）的基础上，该模型提出了“掩码（Masked）语言模型”和“下一句预测模型”. Masked 语言模型通过对语料中 15% 的信息进行遮盖，最大程度地使模型在每 1 个词上都能够学习到全局语境下的表征，从而令 BERT 获得的相关词向量更贴合语境. 具体遮盖方法为 80% 的遮盖信息替换为[MASK]；10% 的遮盖信息替换为任意词；剩余 10% 的遮盖信息保持不变. 同时，BERT 也借鉴了 Skip-thoughts 中的句子预测方法[14]，可以学习句子级别的语义关系：为每个预测样例选择 1 个句子对 A 和 B，让模型预测 A 和 B 是否先后近邻，从而将“下一句预测”问题转化为二分类问题. 其中，50% 的 B 为 A 的下一个句子，标记为“IsNext”；剩余 50% 的 B 为语料库中的 1 个随机句子，标记为“NotNext”. 具体编码过程如下所示.

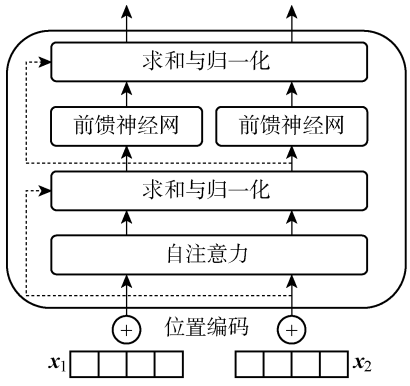

图 2　Transformer 编码器结构

Fig. 2　Structure of Transformer encoder

首先，将输入序列 $\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$ 经过词嵌入（EL）和位置编码（PE）加和后作为 Transformer 编码器的输入：

$$\boldsymbol{X}_e = \mathrm{EL}(\boldsymbol{X}) + \mathrm{PE}(\boldsymbol{X}) \tag{1}$$

式中：$\boldsymbol{X}_e$ 为经过词嵌入和位置编码后的输入序列. 位置编码提供每个字符的位置信息，以便 Transformer 理解句中字词的顺序关系. 词语在句子中的位置不同可能导致语义不同，因此需要对序列中词语的位置进行编码：

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10\,000^{2i/d_{\mathrm{model}}}}\right) \tag{2}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10\,000^{2i/d_{\mathrm{model}}}}\right) \tag{3}$$

式中：pos 为词语在句子中的位置；d_{model} 为 PE 的

维度.

为了提取多重语意含义,输入向量需要经过 1 个多头自注意力机制层:

$$\left.\begin{aligned}\boldsymbol{M}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) &= \mathrm{concat}(\boldsymbol{M}_i)\boldsymbol{W}^0\\ \boldsymbol{M}_i &= \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}})\end{aligned}\right\} \tag{4}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别为查询向量、键向量和值向量;$\boldsymbol{M}_i$ 为单头自注意力机制层;$\boldsymbol{W}^0$ 为权重矩阵;$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$ 和 $\boldsymbol{W}_i^{\boldsymbol{V}}$ 为投影矩阵. 利用缩放点积注意力得到的最终结果为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V} \tag{5}$$

式中:d_{k} 为输入向量的维度. 利用注意力权重对字向量进行加权线性组合,使每个字向量都含有当前句子内所有字向量的信息.

然后,对上一步的输出做一次残差连接($\boldsymbol{X}_1$)和层归一化:

$$\boldsymbol{X}_1 = \boldsymbol{X}_{\mathrm{e}} + \mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{6}$$

$$\mathrm{LayerNorm}(x_i'') = \alpha\frac{x_i''-\mu_{\mathrm{L}}}{\sqrt{\sigma_{\mathrm{L}}^2+\varepsilon}}+\beta \tag{7}$$

式中:x_i'' 为上层输出样本;μ_{L} 和 σ_{L} 分别为均值和标准差;ε 为防止分母为 0 的参数;α 和 β 为弥补归一化过程中损失信息的可训练参数. 残差连接可以避免梯度消失;归一化可以减小数据的偏差,加快训练和收敛的速度.

最后,将经过残差连接和层归一化处理后的信息输入到前馈神经网络中,重复进行一次残差连接和层归一化后输出.

2.2 BiGRU 神经网络

GRU 是 LSTM 的变体. 相比于由 3 个门函数(输入门、遗忘门和输出门)构成的 LSTM,GRU 仅由 2 个门函数构成,即更新门(输入门和遗忘门的结合体,决定过去传递到未来的信息量)和重置门(决定过去信息的被遗忘量). 2 个门控机制能够保存长期序列中的信息,决定哪些信息能够作为门控循环单元的输出. 此外,GRU 具有模型精简、计算速度快、参数少等优势,在小样本数据集上的泛化效果更好. GRU 的具体结构如图 3 所示,表达如下:

$$\left.\begin{aligned}\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}\ \ \boldsymbol{x}_t])\\ \boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}\ \ \boldsymbol{x}_t])\\ \tilde{\boldsymbol{h}}_t &= \tanh(\boldsymbol{W}[\boldsymbol{r}_t\boldsymbol{h}_{t-1}\ \ \boldsymbol{x}_t])\\ \boldsymbol{h}_t &= (1-\boldsymbol{z}_t)\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\tilde{\boldsymbol{h}}_t\end{aligned}\right\} \tag{8}$$

式中:$\boldsymbol{z}$ 和 $\boldsymbol{r}$ 分别为更新门和重置门的输出;$\tilde{\boldsymbol{h}}$ 为记忆内容;σ 为 sigmoid 函数;$\boldsymbol{x}_t$ 为 $\boldsymbol{X}$ 在 t 时刻的分量;

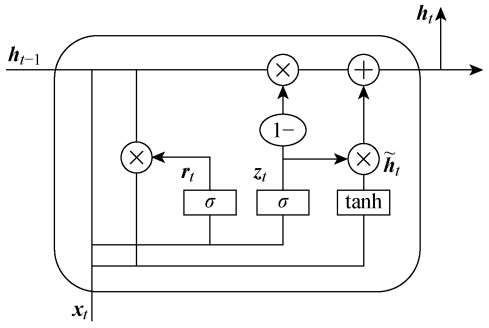

图 3 GRU 结构

Fig. 3 Structure of GRU

$\boldsymbol{W}_z$、$\boldsymbol{W}_r$ 和 $\boldsymbol{W}$ 为权重矩阵.

2.3 注意力机制层

BiGRU 网络在获取语料局部特征上存在不足. 因此,本文利用注意力机制学习句子中任意 2 个字符之间的依赖关系,获取句子的内部结构信息. 注意力机制使命名实体识别模型更专注于挖掘与当前时刻输出相关的输入信息和局部信息. 利用注意力机制对 BiGRU 层输出的特征向量($\boldsymbol{h}_j$)进行权重(a_{tj})分配,计算得到 t 时刻 BiGRU 和注意力机制层共同输出的特征向量($\boldsymbol{c}_t$),并作为最后的输出:

$$\left.\begin{aligned}\boldsymbol{c}_t &= \sum_{j=1}^{T}a_{tj}\boldsymbol{h}_j\\ a_{tj} &= \frac{\exp(e_{tj})}{\sum_{k=1}^{T}\exp(e_{tk})}\\ e_{tj} &= \boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{w}\boldsymbol{c}_{t-1}+\boldsymbol{m}\boldsymbol{h}_j)\end{aligned}\right\} \tag{9}$$

式中:e_{tj} 为对齐模型;$\boldsymbol{v}$、$\boldsymbol{w}$ 和 $\boldsymbol{m}$ 为权重向量.

2.4 CRF 层

BiGRU 层虽然可以学习上下文之间的特征信息,选出最大概率值的标签作为输出,但是不能获取输出标签之间的依赖关系,可能导致 2 个相同标签相互接连. 而 CRF 具有转移特征,能够考虑输出标签之间的顺序性. 因此,选择 CRF 作为 BiGRU 和注意力机制的输出层.

对于每个观察序列,即字符序列 $\boldsymbol{X}'=(x_1',x_2',\cdots,x_n')$,利用线性链条件随机场可以得到 1 个预测标签序列 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$,其预测分数为

$$s(\boldsymbol{X}',\boldsymbol{y}) = \sum_{i=1}^{n}p_{i,y_i}+\sum_{i=0}^{n}A_{y_i,y_{i+1}} \tag{10}$$

式中:p_{i,y_i} 为第 i 个位置标签输出为 y_i 的概率;$A_{y_i,y_{i+1}}$ 为从标签 y_i 转移到 y_{i+1} 的转移概率. 对于每一个 $\boldsymbol{X}'$,得到所有可能的标签序列的分数,则归一化结果和损失函数分别为

$$p(\boldsymbol{y} \mid \boldsymbol{X}') = \frac{\exp(s(\boldsymbol{X}', \boldsymbol{y}))}{\sum_{i=0}^{n} \exp(s(\boldsymbol{X}', \boldsymbol{y}_i))} \tag{11}$$

$$\ln(p(\boldsymbol{y} \mid \boldsymbol{x}')) = s(\boldsymbol{X}', \boldsymbol{y}) - \ln(\sum_{i=0}^{n} \exp(s(\boldsymbol{X}', \boldsymbol{y}_i))) \tag{12}$$

最后，利用维特比(Viterbi)算法[15]得到最佳预测标签序列：

$$\boldsymbol{y}^* = \operatorname{argmax}(s(\boldsymbol{X}, \boldsymbol{y})) \tag{13}$$

Viterbi 算法利用动态规划算法解决 CRF 的预测问题，可以寻找概率最大状态路径.

3　实验与分析

3.1　数据集及标注体系

实验搜集了近 5 年的企业年报，涵盖 2 927 家公司，共 13 129 份. 经过数据清洗和预处理后，按照 6∶2∶2 的比例将其划分为训练集、测试集和开发集. 表 3 为企业年报数据集的详细结构，表 4 为数据集中实体类别个数分布.

表 3　企业年报数据集结构

Tab. 3　Dataset structure of enterprise annual report

项目	字符个数	实体标记字符个数
训练集	384 192	169 263
开发集	128 064	55 294
测试集	133 006	57 426

表 4　实体类别个数分布

Tab. 4　Number distribution of entity categories

实体符号	训练集	开发集	测试集
YEAR	4 884	1 628	1 644
COM	4 182	1 194	1 292
FIN	12 213	4 271	4 297
PNUM	3 421	1 041	1 031
SNUM	3 162	1 016	1 057
PTREND	2 031	629	604
STREND	2 154	627	649

3.2　实验环境和参数设置

在 Python 3.7.3 和 Tensorflow 1.13.1 框架下进行模型的训练和测试. 实验利用 BERT-Base 模型，其含有 12 个 Transformer 层，768 维隐层和 12 头多头注意力机制. GRU 网络的隐层设为 128 维. 注意力机制层设置为 50 维，最大序列长度设置为 256. 优化函数采用 Adam，学习率设置为 5×10^{-5}，dropout 层设置为 0.5.

3.3　评估标准

实验利用精确率(P)、召回率(R)和 F_1 值共 3 个指标评价 7 大类实体的命名实体识别效果，3 个评价指标的计算方法如下：

$$P = a/b \tag{14}$$

$$R = a/c \tag{15}$$

$$F_1 = 2PR/(P+R) \tag{16}$$

式中：a 为正确识别实体数；b 为识别实体总数；c 为所有实体总数.

4　实验结果与分析

BERT-BiGRU-Attention-CRF 模型对不同实体的识别效果如表 5 所示. 其中“年份”“企业名称”“财务术语”“利润类/收入类数值”实体有较高的 P、R 和 F_1 值. 模型对“利润类/收入类数值同上年变化趋势”实体的识别性能相对较差，主要是由于该类实体表达形式较复杂，包括纯文字表达、文字和数字组合表达等，且描述变化趋势的文字表达形式多样. 对此，可以通过深入的划分实体、融合词典特征和改进模型等方式，令实体学习更多语义特征.

为了验证 BERT-BiGRU-Attention-CRF 模型在企业年报命名实体识别中的优异性，在同一数据集上，分别对 CRF、BiGRU-CRF、BiGRU-Attention-CRF 和 BERT-BiGRU-CRF 模型进行实验，对比结果如表 6 所示. 此外，利用雷达图显示不同实体

表 5　不同实体识别效果

Tab. 5　Recognition effect of different entities　%

实体符号	P	R	F_1
YEAR	98.66	98.66	98.66
COM	96.59	93.41	94.97
FIN	95.93	93.07	94.48
PNUM	92.97	89.81	91.37
SNUM	95.83	92.00	93.88
PTREND	91.46	79.44	85.03
STREND	93.05	86.73	89.73

表 6　不同模型实验结果

Tab. 6　Experimental result of different models　%

模型	P	R	F_1
CRF	74.52	71.20	72.82
BiGRU-CRF	86.69	83.61	85.12
BiGRU-Attention-CRF	88.75	86.28	87.50
BERT-BiGRU-CRF	94.19	90.07	92.08
BERT-BiGRU-Attention-CRF	95.45	91.99	93.69

在不同模型上的 F_1 值，如图 4 所示. 由图 4 可知，BERT-BiGRU-Attention-CRF 模型在 7 大类实体上的 F_1 值都处于较高水平，说明该模型在企业年报领域的识别性能高于其他模型.

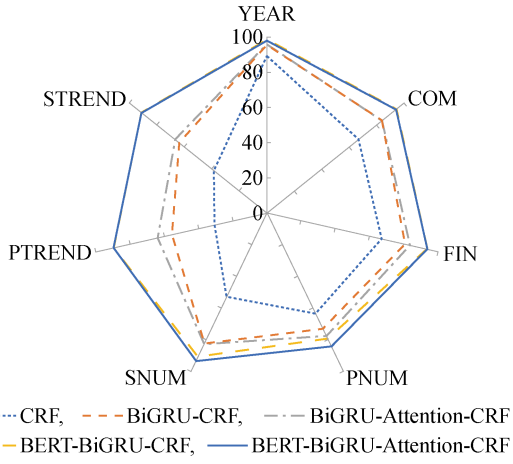

图 4 不同实体在不同模型上的 F_1 值(%)

Fig. 4 F_1 values of different entities in different models (%)

不同模型结合不同实体的具体分析如下：

(1) CRF 模型是基于统计的命名实体识别方法，由于 CRF 是在分词的基础上通过设置特征模板获取语料的特征，所以对“企业名称”“财务术语”“利润类/收入类数值”和“利润类/收入类数值同上年变化趋势”这 4 类属于未登录词的实体识别效果较差，其 F_1 值均在 68%以下.

(2) 相比于 CRF 模型，BiGRU-CRF 模型整体的 F_1 值提高了 12.3%，且对未登录词实体的边界划分更准确. 这是因为未登录词实体的构成较复杂、词长较长，CRF 特征模板只能在有限的窗口范围内进行提取，而 BiGRU 网络可以更好地利用上下文的语义特征，如更善于区分“净利润”和“归属于母公司的净利润”这类易混淆词语、识别出更多完整的企业名称和简称.

(3) 相比于 BiGRU-CRF 模型，BiGRU-Attention-CRF 模型的 F_1 值提高了 2.38%. 句子中不同的字词和上下文的关联程度不同，而注意力机制可以关注更多的局部特征，特别是和当前输出有关联的信息，如识别句“实现净利润 13 亿元”中的“利润类数值”实体，词语“净利润”与实体的关联程度大于词语“实现”，则注意力机制会更关注“净利润”和实体之间的关系.

(4) 相比于 BiGRU-CRF 模型，BERT-BiGRU-CRF 模型的 F_1 值提高了 6.96%；相比于 BiGRU-Attention-CRF 模型，BERT-BiGRU-Attention-CRF 模型的 F_1 值提高了 6.18%，具体反映为“收入类数值”和“利润类数值同上年变化趋势”2 类实体的 F_1 值分别提高 10.38%和 25.31%. 这 2 类实体和上下文之间的关联较强，且表达方式较灵活，如在字词级别方面，“数值”实体中单位的表示方式有元、万元、亿元等；在句子级别方面，“数值同上年变化趋势”实体有文字-数字结合(涨幅/增长/下降＋百分比)和纯文字描述(创下新高、扭亏为盈)共 2 种表达方式. 此外，融入 BERT 模型的企业年报命名识别方法更能够结合语义找到数值和财务术语的映射关系，尤其适用于同时包含 2 个数值的句字，如“营业收入和主营业务收入分别为 13 万元和 10 万元”. 综上可知，BERT 通过在大型语料库上学习获得更多语义特征，可以对企业年报这一小规模语料库进行特征补足；其利用双向 Transformer 结构进行基于上下文语境的深度双向语义理解，提高特征抽取的能力和边界不明显且表述灵活实体的识别效果. 此外，BERT 能够学习字符级、词级和句子级关系特征，可以更全面地理解句子语义.

5 结语

企业年报命名实体识别为企业评价系统的建设提供了基本企业信息和经营情况的数据支撑. 本文提出了 BERT-BiGRU-Attention-CRF 模型. 在基础模型 BiGRU-CRF 上引入 BERT 预训练语言模型，以获得与上下文有关联的双向特征表示，更加深刻地理解语义，克服了企业年报语料库规模小、实体专业性和映射关系强的问题. 然后，采用注意力机制改进 BiGRU-CRF 模型，使模型可以选择性地关注重要信息，提高信息的有效关注率. 自建企业年报语料库的识别结果表明：BERT-BiGRU-Attention-CRF 模型能够较好地识别企业年报中的实体，可以达到 95.45%的精确率和 91.99%的召回率以及 93.68%的 F_1 值，能够满足应用需求. 在后续研究中，将扩大语料库规模，进一步完善并规范企业年报的实体标注，提取更多有价值的实体，并在保证性能的基础上，对模型结构进行简化.

参考文献：

[1] ZHENG S C, WANG F, BAO H Y, *et al*. Joint extraction of entities and relations based on a novel tagging scheme[C]// **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers)**. Stroudsburg, PA, USA: ACL, 2017: 1227-1236.

[2] 吴文涛，李培峰，朱巧明. 基于混合神经网络的实体和事件联合抽取方法[J]. **中文信息学报**，2019，33(8)：77-83.
WU Wentao，LI Peifeng，ZHU Qiaoming. Joint extraction of entities and events by a hybrid neural network[J]. **Journal of Chinese Information Processing**，2019，33(8)：77-83.

[3] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[OL]. (2015-08-09)[2019-11-20]. https：// arxiv. org/abs/1508.01991v1.

[4] CHIU J P C，NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. **Transactions of the Association for Computational Linguistics**，2016，4(1)：357-370.

[5] MA X Z，HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[C]// **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.** (**Volume 1：Long Papers**). Stroudsburg，PA，USA：ACL，2016：1064-1074.

[6] CHO K，VAN M B，GULCEHRE C，*et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing** (**EMNLP**). Stroudsburg，PA，USA：ACL，2014：1724-1734.

[7] 王洁，张瑞东，吴晨生. 基于 GRU 的命名实体识别方法[J]. **计算机系统应用**，2018，27(9)：18-24.
WANG Jie，ZHANG Ruidong，WU Chensheng. Named entity recognition method based on GRU[J]. **Computer Systems & Applications**，2018，27(9)：18-24.

[8] BHARADWAJ A，MORTENSEN D，DYER C，*et al*. Phonologically aware neural model for named entity recognition in low resource transfer settings [C]// **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.** Stroudsburg，PA，USA：ACL，2016：1462-1472.

[9] CAO P F，CHEN Y B，LIU K，*et al*. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]// **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.** Stroudsburg，PA，USA：ACL，2018：182-192.

[10] VASWANI A，SHAZEER N，PARMAR N，*et al*. Attention is all you need[C]// **Proceedings of the 31st Conference on Neural Information Processing Systems.** Long Beach，CA，USA：NIPS，2017：5998-6008.

[11] DEVLIN J，CHANG M W，LEE K，*et al*. BERT：Pre-training of deep bidirectional transformers for language understanding[OL]. (2019-05-24)[2019-11-25]. https：//arxiv.org/abs/1810.04805.

[12] WITTEN I H，PAYNTER G W，FRANK E，*et al*. KEA：Practical automatic keyphrase extraction[C]// **Proceedings of the Fourth ACM Conference on Digital Libraries.** New York，USA：ACM，1999：254-255.

[13] KANDOLA E J，HOFMANN T，POGGIO T，*et al*. A neural probabilistic language model[J]. **Studies in Fuzziness and Soft Computing**，2006，194：137-186.

[14] KIROS R，ZHU Y K，SALAKHUTDINOV R R，*et al*. Skip-thought vectors[C]// **Proceedings of the 29th Conference on Neural Information Processing Systems.** Montreal，Canada：NIPS，2015：3294-3302.

[15] STRUBELL E，VERGA P，BELANGER D，*et al*. Fast and accurate entity recognition with iterated dilated convolutions[C]// **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.** Stroudsburg，PA，USA：ACL，2017：2670-2680.

（本文编辑：孙伟）