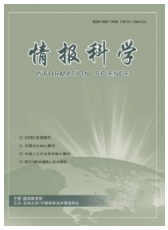

情报科学
Information Science
ISSN 1007-7634,CN 22-1264/G2

《情报科学》网络首发论文

题目：金融领域文本序列标注与实体关系联合抽取研究
作者：唐晓波，刘志源
收稿日期：2020-04-17
网络首发日期：2020-10-23
引用格式：唐晓波，刘志源．金融领域文本序列标注与实体关系联合抽取研究[J/OL]．情报科学. https://kns.cnki.net/kcms/detail/22.1264.G2.20201022.1502.004.html

网络首发时间：2020-10-23 15:40:11
网络首发地址：https://kns.cnki.net/kcms/detail/22.1264.G2.20201022.1502.004.html

金融领域文本序列标注与实体关系联合抽取研究

唐晓波[1,2]，刘志源[1]

（1.武汉大学 信息管理学院，湖北 武汉 430072；2.武汉大学 信息资源研究中心，湖北 武汉 430072）

摘　要：【目的/意义】金融领域实体关系抽取是构造金融知识库的基础，对金融领域的文本信息利用具有重要作用。本文提出金融领域实体关系联合抽取模型，增加了对金融文本复杂重叠关系的识别，可以有效避免传统的流水线模型中识别错误在不同任务之间的传递。【方法/过程】本文构建了高质量金融文本语料，提出一种新的序列标注模式和实体关系匹配规则，在预训练语言模型BERT（Bidirectional Encoder Representations from Transformers）的基础上结合双向门控循环单元BiGRU（Bidirectional Gated Recurrent Units）与条件随机场CRF（Conditional Random Field）构建了端到端的序列标注模型，实现了实体关系的联合抽取。【结果/结论】针对金融领域文本数据进行实验，实验结果表明本文提出的联合抽取模型在关系抽取以及重叠关系抽取上的F1值分别达到了0.627和0.543，初步验证了中文语境下本文模型对金融领域实体关系抽取的有效性。

关键词：关系抽取；联合抽取；文本序列标注；BERT；BiGRU

Research on Text Sequence Tagging and Joint Extraction of Entity and Relation in Financial Field

TANG Xiao-bo[1,2]，*LIU Zhi-yuan*[1]

(1. School of Information Management, Wuhan University, Wuhan 430072*, China;*
2. Center for Studies of Information Resources, Wuhan University, Wuhan 430072*, China)*

Abstract: [Purpose/Significance] Entity relation extraction in financial field is the basis of constructing financial knowledge base and plays an important role in the utilization of text information in financial field.This paper proposes a joint extraction model of entity relations in the financial field, which increases the recognition of complex overlapping relationships of financial texts, and can effectively avoid the erroneous delivery between different tasks in the traditional pipeline model. 【Method/process】This paper constructs a high-quality financial text corpus, proposes a new tagging scheme and entity relation matching rule, constructs an end-to-end sequence annotation model based on BERT and combines BiGRU and CRF, and realizes the joint extraction of entity relation. [Result/Conclusion] Experiment with text data in the financial field, the experimental results show that the F1-score of the relationship extraction and overlapping relationship extraction of the model proposed in this paper reach 0.627 and 0.543 respectively, which preliminarily verifies the validity of the joint extraction model proposed in this paper for entity relationship extraction in the Chinese financial field.

Keywords: relation extraction; joint extraction; text sequence tagging; BERT; BiGRU

1 引　言

金融在人类社会中扮演着十分重要的角色，金融活动产生了海量有用的信息资源，然而其中大部分都是难以直接利用的非结构化信息。信息抽取技术可以有效的将非结构化信息转化为结构化信息以便更好的利用，具体包括命名实体识别、实体关系抽取、事件抽取等研究方向。对金融领域文本进行信息抽取是构建金融知识库的基础，对金融大数据的利用具有重要的意义。

收稿日期：2020-04-17
基金项目：国家自然科学基金项目“基于文本和Web语义分析的智能咨询服务研究”（71673209）
作者简介：唐晓波（1962-），男，教授，博士生导师，主要从事知识组织与情报研究.

实体关系抽取是研究者们对高效自动地从海量信息中获取知识所做出的探索，它是知识图谱构建、语义理解等任务的重要前期工作之一。比如，给定例句："湖北省的省会是武汉市。"实体关系抽取最终将抽取得到(湖北省，省会城市，武汉市)这一实体与关系的知识三元组。对于传统流水线方式实体关系抽取，将命名实体识别与实体关系抽取两个任务分离进行，各任务可灵活调整，但后续的关系抽取依赖于命名实体识别的结果，如果实体识别错误则不会有正确的关系抽取结果。对于部分现有的联合抽取模型，需要大量的复杂的特征工程，需要使用词本身以及词与词之间的依存关系等。而部分基于神经网络的端到端关系抽取模型所使用的词嵌入方法无法有效的对一词多义情况进行编码，从而也会影响抽取效果。

鉴于此，本文提出基于BERT[1]的金融领域实体关系联合抽取模型。在预先给定关系类型情况下，提出一种考虑到句中重叠关系抽取的序列标注方式和三元组抽取规则，将实体与关系抽取两个子任务转换为序列标注问题，在此基础上构建了端到端的BERT-BiGRU-CRF序列标注模型。本文提出的联合抽取模型增强了对字的语义表征，不再需要复杂的特征工程，避免了流水线方式的识别错误传递。

2 相关研究

实体关系抽取旨在从文本中挖掘出所需要的实体关系信息，由于文档级别的关系抽取数据标注难度大，缺少相应的人工大规模标注数据集，因此当前大部分的关系抽取工作聚焦于句子级的文本粒度进行。作为信息抽取领域的重要研究研究课题，实体关系抽取是知识图谱构建、知识库构建等任务的前期工作，为其它自然语言处理应用提供了技术支持。按照实体关系抽取工作的处理方式不同，可分为流水线方式(Pipelined Method)和联合学习方式(Joint Learning Method)。

流水线方式将实体关系抽取分离为两个任务，即先进行命名实体识别(Named Entity Recognition, NER)再对实体之间进行关系分类(Relation Classification, RC)。命名实体识别任务最早是由Grishman等人于1996年在第六届MUC(Message Understanding Conference)会议上提出的[2]，经典解决方案主要包括基于规则的方法与基于统计机器学习的方法[3]。早期Cucerzan等人[4]提出了一种具有语言独立性，结合了上下文语境的命名实体识别规则自动生成的自举算法，然而制定规则的方法虽然可以在某些情境下取得不错的识别效果，但是不同领域之间的规则往往不具有可移植性；随着计算能力的提升以及词的向量表示等方面获得突破，人们逐步将目光转移到了机器学习的方法上，zhou等人[5]利用隐马尔可夫模型(Hidden Markov Model, HMM)建立了命名实体识别系统取得了优于利用人工规则识别的效果，Duan等人[6]基于条件随机场结合汉字词性、前缀等特征，对人物、地点和机构进行了识别；随着深度学习在机器学习领域的兴起以及词向量技术的发展，将学者从繁重的人工特征工程中解放了出来，Lample等人[7]提出了基于双向长短时记忆网络(Bidirectional Long Short-Term Memory, Bi-LSTM)与条件随机场的结构用于命名实体识别，Strubell等人[8]采用了迭代膨胀卷积神经网络(Iterated Dilated Convolutional Neural Network，IDCNN)进行命名实体识别。

在关系分类子任务上，基于特征向量、核函数的机器学习方法有广泛的应用[9]，车万翔等人[10]在选取实体左右两个词作为特征时，利用支持向量机(Support Vector Machine, SVM)与Winnow算法进行关系抽取有较好的效果，Zhang等人[11]在改进的卷积树核函数基础上提出了一种复合核函数来进行关系抽取，并在ACE语料上取得了最佳的结果；随着深度学习的兴起，实体关系抽取有了新的探索方向，Liu等人[12]最早提出将卷积神经网络(Convolutional Neural Networks, CNN)用于关系分类，比非神经网络方法的效果有了显著提升，Zhou等人[13]提出了基于注意力的双向长短时记忆网络，在SemEval-2010关系分类任务中取得了较优结果，Cai等人[14]提出了双向递归卷积神经网络(Bidirectional Recursive Convolutional Neural Network, BRCNN)模型，并利用实体的最短路径依赖的信息进行实体关系分类。流水线方式进行的关系抽取将两个子任务分离，可以使框架更加灵活，虽然每个子任务都分别取得了良好的成果，但分离框架忽略了任务之间的相关性，且会导致错误在任务之间的传递[15]。

联合学习方式的实体关系抽取将两个任务并入同一个模型框架内，Yu等人[16]提出了一种概率图模型用于同时优化相关子任务，Li等人[17]利用了结构化感知器同时提取实体和关系，但是以上方法都需要依赖复杂的特征工程；Zheng等人[15]将实体关系联合抽取转化为序列标注问题，并结合端到端模型获得了较好的实验结果，但是对于复杂句中一个实体同时属于多个关系的重叠关系情况则无法有效识别抽取；曹明宇等人[18]为了缓解生物医学文本中大量重叠的实体关系问题，改进了Zheng等人[15]的标注方式，引入了重叠标签并在药物实体关系抽取上体现了更高的性能，但不能覆盖句中所有复杂的重叠关系；上述论文直接使用词向量特征，摆脱了复杂的特征工程，但是对于使用Word2Vec[19]，Glove[20]等词向量表示方法的神经网络模型来说，无法对汉字或英文单词的多义性进行有效表征。

为了解决流水线方式的信息抽取存在的误差传递等问题，提升对金融领域大量存在的复杂重叠实体关系的识别，优化传统向量表示方法无法对字符多义性有效表征的问题；本文使用了一种改进标注方式将实体关系抽取转换为序列标注问题，利用预训练语言模型BERT的强大语义表示能力，BiGRU对上下文的理解能力和CRF对状态序列之间联系的学习能力，构建了端到端的BERT-BiGRU-CRF序列标注模型；再根据标注结果以特定的匹配规则完成实体关系抽取。本文主要有以下贡献：引入预训练语言模型BERT以增强对汉字多义性的向量表示；改进标注模式，在一定程度上缓解复杂重叠关系难以有效标注的问题；建立联合抽取模型

用于同时对实体和关系进行抽取以减少将不同任务分离的流水线模式具有的缺陷。

3 基于BERT的金融领域实体关系联合抽取模型

基于BERT的金融领域实体关系联合抽取模型由三部分组成，分别是文本数据预处理，BERT-BiGRU-CRF序列标注模型以及实体关系三元组抽取，如图1所示。

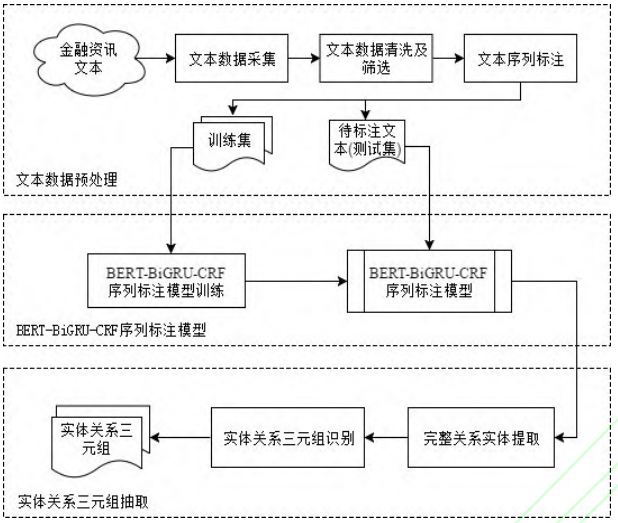

图1 基于BERT的金融领域实体关系联合抽取模型

3.1 文本数据预处理

针对金融领域语料数据的预处理主要包括以下几点：

(1)数据采集：从东方财富网采集上市公司资讯信息。由于全篇资讯的主要信息存在于概述当中，为便于后续数据清洗工作，本文选择采集资讯的概要部分。

(2)数据清洗及筛选：包括过滤无关文本以及对可标注文本的筛选。东方财务网的资讯充斥大量不存在实体关系的数值说明型文本，例如每日融资融券余额、涨幅通报等类型的文本，且实体数低于两个或者只存在不属于已定义五种关系类别的实体关系的数据需要进行剔除。因此，在初步过滤特征较明显的无关文本后，为保证文本质量，以句子为单位对文本进行人工筛选。

(3)序列标注：清洗及筛选之后的文本数据进行人工标注工作。

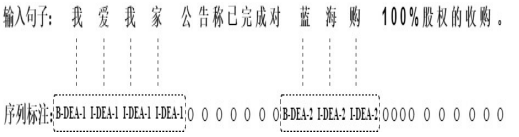

图2 一般关系标注示例

一般关系标注方案如图2所示。句中，每一个字都被赋予一个标签，本文提出的标签包含最多不超过3部分，依次分别是实体边界、关系类别和实体角色。

对于实体边界标签，本文采用“BIO”标注方式来表示单字在实体中的位置信息，B表示此元素在实体的头部，I表示此元素在实体的中部或尾部，O则表示该元素不属于任何实体。

实体关系标签由预先定义的关系类别来确定，在王树伟[21]与孔兵[22]的金融实体关系分类基础上结合本文采集金融领域资讯语料，本文如表1所示将主要关系分为5类，在实际标注中取其英文单词前三位大写作为对应实体关系标签。

实体角色标签由数字“1”和“2”表示，代表实体在所抽取的三元组中的先后顺序，即(实体1，关系类别，实体2)，在标注过程中可通过在三元组中的实体先后顺序赋予关系特殊的含义，在本文抽出的知识三元组中，对不同关系的实体角色标签要求如下：买卖关系中，用实体角色标签“1”来标注买卖动作的主动发出者，否则标为“2”；股权关系中，用实体角色标签“1”来标注股东，被控股的企业组织标为“2”；合作关系较为特殊，按照句中先后出现顺序进行标注；角色关系当中，用实体角色标签“1”来标注人物角色，企业组织标为“2”；处罚关系中，用实体角色标签“1”来标注处罚相关动作的发出者，接受者标为“2”。

对于重叠关系，本文引入了一种新的标签类型，由实体边界标签与重叠标签“OVE”构成，重叠关系实体的标签不再需要标注实体角色，它的实体角色标签由与它配对的实体确定，如配对实体的实体角色标签为“2”，则重叠关系实体的实体角色标签默认为“1”，反之亦然。如图3所示，例句中包含两组关系分别是“熊猫金控”旗下控股子公司“银湖网”，二者包含了股权关系，以及“公安”对“银湖网”立案，包含了处罚

表1 金融领域实体关系类型

关系类型	说明	例句
买卖(Deal)	金融资讯报道中常见的商业行为，发出者可以是人或者公司，包括公司之间的股份、商品等交易。	近日，我爱我家控股集团以出资方式完成对美住网51%的股权收购。
股权(Stock Equity)	在企业网络中普遍存在复杂的股权关系，常以控股子公司，控股股东等表述形式出现。	2019年9月9日，京汉股份发布公告称，公司控股子公司南京空港领航发展有限公司负责建设和运营南京空港保税物流中心(B型)。
合作(Cooperate)	企业组织之间商业行为，通常以签订某种合约体现，包括共同出资、合营等合作行为。	2019年11月1日，南宁市人民政府与绿地控股(600606)战略合作协议签约暨中国(广西)自由贸易试验区南宁片区绿地东盟总部项目揭牌。
角色(Role)	代表了人物与组织公司间具有某种关系，如公司董事长、实际控制人、总经理等。	10月9日晚间，世联行发布公告称，世联行董事会同意董事长陈劲松代行第五届董事会秘书一职，直至公司正式聘任新的董事会秘书。
处罚(Punish)	在金融资讯中较常见并能对上市公司股价造成较大影响的处罚相关行为，例如证监会对企业或个人罚款，经侦部门对企业或管理层立案调查等。	中国银保监会金华监管分局近期对平安银行义乌分行开具了罚单，该行将信贷资金用于支付购房首付款等，被罚款人民币50万元。

关系。实体“银湖网”同时属于两种关系，发生了关系的重叠，具有重叠标签的实体可与除重叠外任何关系类别的实体配对。

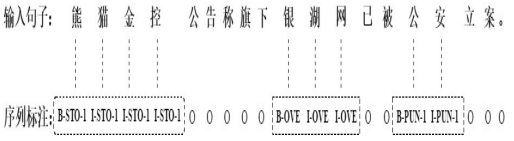

图 3 重叠关系标注示例

实际标签数量分3类共计23种，如表2所示。

表 2 金融实体关系标签类型

标签种类	示例	数量
非实体	O	1
一般关系实体	B-DEA-1、I-DEA-1、B-STO-1、I-STO-1、B-COO-1、I-COO-1、B-ROL-1、I-ROL-1、B-PUN-1、I-PUN-1、B-DEA-2、I-DEA-2、B-STO-2、I-STO-2、B-COO-2、I-COO-2、B-ROL-2、I-ROL-2、B-PUN-2、I-PUN-2	20
重叠关系实体	B-OVE、I-OVE	2

3.2 BERT-BiGRU-CRF序列标注模型

BERT-BiGRU-CRF序列标注模型的整体结构如图4所示，整个模型依次由BERT层、双向GRU和CRF层共三部分组成。句子输入BERT预训练语言模型层，获得每个字的基于上下文计算的向量表示后，将字的向量输入BiGRU层得到每个字对于各标签的非归一化概率分布，将其作为CRF层的输入最终得到考虑标签之间依赖关系的全局最优标签序列。

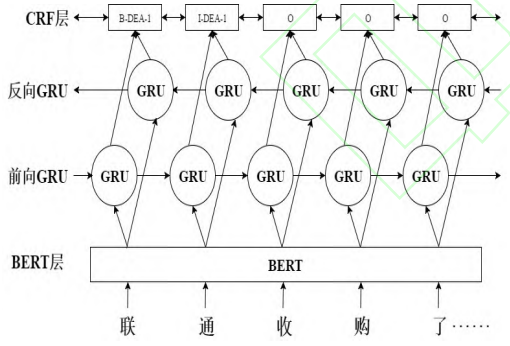

图 4 BERT-BiGRU-CRF序列标注模型

3.2.1 BERT层

BERT模型由Devlin等人[1]于2018年提出，具体结构如图5所示，它是基于Transformer的双向编码表征模型，与OpenAI GPT[23]和ELMo[24]等语言模型最大的区别是BERT使用了双向的Transformer结构，在预训练阶段使用了遮蔽语言模型(Masked Language Model，MLM)，以及下一句预测(Next Sentence Prediction)两个任务进行联合训练来使得表征能融合上下文。遮蔽语言模型是作者为了训练深度双向表示所采用的一种方法，作者随机遮盖每个序列中15%的单词，对被遮盖的词进行预测，而被随机选择进行遮盖的单词中有80%用【MASK】标记对单词进行替换，10%用随机单词替换，10%的单词保持不变，通过以上做法，可以使Transformer保持对每个输入token的分布式上下文表示，且不会损害模型的语言理解能力。下一句预测是一个二分类预测，对每个预测选择一个句子对A和B，B有50%的概率是A的下一个句子，50%的概率是语料库当中随机的一个句子，使得语言模型能理解两个句子之间的关系。

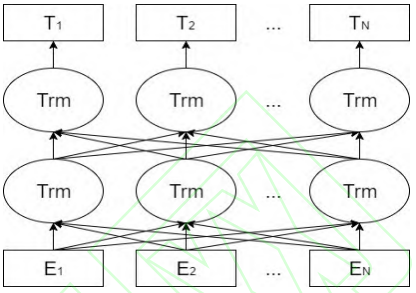

图 5 BERT预训练语言模型

Transformer不同于传统的循环神经网络(Recurrent Neural Network，RNN)结构，它基于自注意力机制(Self-Attention)来进行并行计算，Transformer编码单元如图6所示。

在输入部分，Transformer额外添加了一个位置编码(Positional Encoding)向量，维度与输入的单词embedding维度一致，以加入相对位置信息，pos指当前词在句中的位置，d_{model}表示位置向量中每个值的维度，在偶数位使用正弦编码，奇数位使用余弦编码。计算公式如下：

$$PE(pos,2i)=\sin(pos/10000^{2i/d_{model}}) \quad (1)$$

$$PE(pos,2i+1)=\cos(pos/10000^{2i/d_{model}}) \quad (2)$$

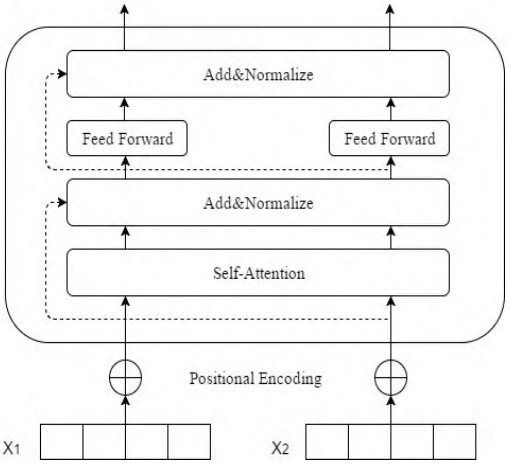

图6 Transformer编码单元

编码单元当中最重要的是自注意力模块，自注意力的计算如公式(3)所示，Q,K 和 V 均为输入的字向量矩阵，全称是Query、Key和Value，由embedding向量与一个随机初始化的矩阵相乘得到；QK^T 代表的是Query与Key做点乘得到的分值，决定了在某个位置编码一个词时对句子输入的其他部分关注程度；d_k 为输入向量维度，将点乘的结果除以输入向量维度的开方进行缩放以防止分值随着向量维度增大；最后将softmax归一化的值与 V 矩阵相乘；这种通过Query与Key的相似程度来确定Value权重分布的方法被称为Scaled

Dot-Product Attention，这样每个词都能蕴含对句中其他词的关系，有了更为全局的表达。

$$Attention(Q,K,V)=soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

除此之外 Transformer 还引入了多头（Multi-Head）机制。假设头数为 n，则初始化 n 组 Q,K 和 V，分别通过不同参数矩阵（W_i^Q,W_i^K,W_i^V）映射后再接入自注意力模块计算 Attention 的值，最后将结果拼接送入一个全连接层，W^o 为附加权重矩阵，如公式（4）（5）所示。多头机制增加了模型捕获不同位置信息的能力，且因进行映射时不共享权值，映射后的子空间所涵盖的信息不同，拼接后向量可具有更多的信息。

$$MultiHead(Q,K,V)=Concat(head_1,...,head_n) \tag{4}$$

$$\text{head}_i=Attention(QW_i^Q,KW_i^K,VW_i^V) \tag{5}$$

Transformer 中还加入了了残差网络和层归一化以改善多层堆叠时可能产生的退化问题，前馈神经网络层（Feed Forward）采用了全连接层加 ReLU 函数实现，如公式（6）（7）所示：

$$LN(x_i)=\alpha\times\frac{x_i-\mu_L}{\sigma_L^2+\varepsilon}+\beta \tag{6}$$

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{7}$$

BERT 作为基于 Transformer 的双向编码语言模型相比较其他语言模型而言，可以充分利用词语上下文的信息，从而使得单词或汉字获得更好的向量表示。

3.2.2 BiGRU 层

GUR 是循环神经网络中的一种。循环神经网络不同之处在于隐层之间也建立了权连接，使得当前隐层的信息可以传递到下个节点，序列中的节点能够考虑到前文的影响，从而在处理序列数据上具有优势。

循环神经网络理论上可以处理任意长度的序列问题，但实际应用容易出现序列过长导致梯度消失的问题从而导致长期依赖难以学习[25]。Cho 等人[26]提出的 GRU 解决了长期记忆和反向传播中的梯度问题，不同于 1997 年提出的 LSTM（Long Short-Term Memory）[27]，GRU 简化了结构，减少了参数，将遗忘门和输入门合为更新门并混合了细胞状态和隐藏状态，单元结构如图 7 所示。

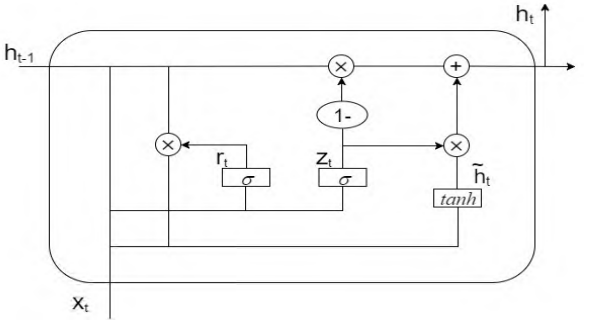

图 7 GRU 单元结构

GRU 共有两个门，分别为更新门 z_t 和重置门 r_t，更新门决定了前一时刻状态信息能带入当前状态中的程度，它的取值越大代表前一时刻进入当前状态的信息越多，重置门决定了前一时刻的状态信息有多少需要遗忘，它的值越小代表遗忘得越多。计算公式如下：

$$z_t=\sigma(W_z x_t+U_z h_{t-1}) \tag{8}$$

$$r_t=\sigma(W_r x_t+U_r h_{t-1}) \tag{9}$$

$$\tilde{h}_t=\tanh(Wx_t+U(r_t\odot h_{t-1})) \tag{10}$$

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\tilde{h}_t \tag{11}$$

其中 σ 是 sigmoid 函数，所有 W 、U 均为权重参数，$\odot$ 代表 Hadamard 乘积，h_t 是输出，也是当前时刻的隐藏状态，h_{t-1} 表示上一时刻隐藏状态。

BiGRU 分为前向 GRU 与反向 GRU，同 BiLSTM[28]一样，避免了单向结构只能考虑到历史信息的缺陷，使模型能够更好利用上下文信息，前向 GRU 用以捕捉上文信息，反向 GRU 用以捕捉下文信息，最终将正向 GRU 与反向 GRU 的输出进行拼接，作为 CRF 的输入。GRU 结构较 LSTM 简单参数较少，训练速度更快，在序列标注任务上有广泛的利用[25]。

3.2.3 CRF 层

文本经过 BERT 与 BiGRU 层后只能得到文本序列与标签之间的关系，无法考虑标签与标签之间的关系。本文标注模式中，当前字的标签与上下文的标签具有联系，例如在同一个实体中每个字的关系标签必须一致、实体边界标签中的 I 标签只能在 B 标签之后等。因此在 BiGRU 之上增加 CRF 层用以获得全局最优标签序列。

对于给定输入句子序列 $x=\{x_1,x_2,x_3,...,x_n\}$ 经 BiGRU 层获得输出，矩阵 P 为 BiGRU 层输出结果，其大小为 $n\times m$，n 为汉字个数，m 为标签种类。$P_{i,j}$ 代表句子中第 i 个字的第 j 个标签的概率。对于句子序列 x 对应的标签序列 $y=\{y_1,y_2,y_3,...,y_n\}$，CRF 定义评估分数如下：

$$s(x,y)=\sum_{i=1}^{n}(W_{y_{i-1},y_i}+P_{i,y_i}) \tag{12}$$

其中 W 是转移矩阵，$W_{i,j}$ 代表由标签 i 转移到标签 j 的概率。

3.3 实体关系三元组抽取

对于序列标注结果，需按一定配对规则将其转换为知识三元组。首先通过实体边界标签获取完整实体，对于某个具有一般关系标签的实体，就近寻找可配对实体，由于重叠关系实体可以与任何实体配对，因此具有重叠关系标签的实体不得由自身出发主动和其他实体配对。每一组以句为单位的标签序列的实体关系三元组抽取步骤如下：

（1）通过实体边界、实体关系以及实体角色标签得到完整关系实体，并得到其对应的关系类型、实体角色以及在句中的位置，当且仅当实体边界标签符合“B”为头部“I”为非头部，实体中每一个字的关系标签和角色标签一致方可提取。

（2）对于每一个提取完毕的普通关系实体，分别向前和先后查找可配对实体。可配对实体需满足以下两条件之一：①可配对实体的实体关系标签与该实体一致，且实体角色标签与该实体不同。②可配对实体为重叠关系实体。

（3）得到前向和后向查找到的可配对实体后，分别判断

该实体与可配对实体间的距离，将距离最近的一组实体配对形成实体关系三元组，若匹配成功的两个实体均为普通关系实体则两实体均不再参与后续的配对，若其中含有重叠关系实体，则仅使重叠关系实体继续参与后续的配对。重复步骤(2)(3)，直到句子中的所有非重叠关系实体均处理完成。

普通关系三元组提取如图8所示，通过实体边界、实体关系以及实体角色标签提取实体一“我爱我家”以及实体二“蓝海购”，以实体一为基准，向其之前和之后进行可配对实体查找，由于实体二的实体关系与实体一相同，实体角色标签不同，且在该句中只存在一个可配对实体，因此匹配结果为(我爱我家，买卖，蓝海购)，两者均不再参与匹配，该句实体关系三元组抽取完毕。

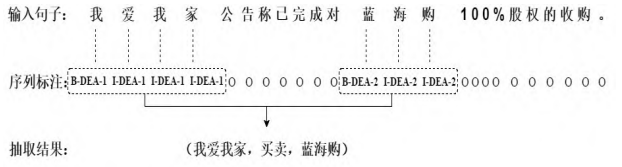

图8 普通关系三元组抽取示例

重叠关系三元组提取如图9所示，通过实体边界、实体关系以及实体角色标签提取实体一“熊猫金控”，实体二“银湖网”以及实体三“公安”。按实体的顺序，以实体一为基准向前和向后查找可配对实体，可以得到距离最近可配对实体“银湖网”，得到第一个配对结果为(熊猫金控，股权，银湖网)，由于“银湖网”为重叠关系实体，可继续参与后续的配对。重叠关系实体不主动参与匹配，因此略过实体二“银湖网”，以实体三为基准向前和向后查找可配对实体，同样可以得到距离最近可配对实体“银湖网”，得到第二个配对结果(公安，处罚，银湖网)，该句所有非重叠关系实体均处理完毕，实体关系三元组抽取完毕。

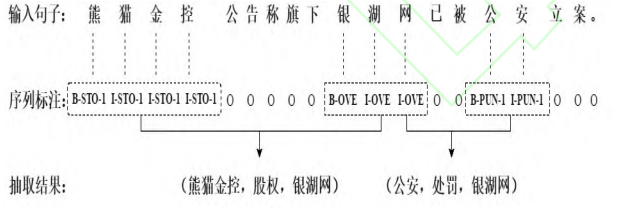

图9 重叠关系三元组抽取示例

存在某些复杂句形式，如图10所示，无法通过本文所述的标注模式进行有效标注和关系抽取。

例句： 11月8日晚，万达信息发布公告称，截至当日，中国人寿及其一致行动人中国人寿保险(集团)公司、中国人寿资产管理有限公司合计持有公司20632.37万股股份，已超过万豪投资及其实际控制人史一兵合计持有的20609.74万股股份，成为公司第一大股东。

图10 无法有效标注例句

例句按本文的标注模式，“万达信息”作为重叠关系实体，“中国人寿”、“中国人寿资产管理有限公司”等实体均为“万达信息”的股东，但是“万豪投资”这一实体的实际控制人是“史一兵”，两者具有角色关系。若将“万豪投资”实体标注为股权关系，则会因此丢失其自身具有的角色关系，反之则会丢失其与万达信息的股权关系；若将“万豪投资”实体标注为重叠关系实体，则按照本文提出的匹配规则会导致“万豪投资”错误的与“中国人寿资产管理有限公司”配对成功，故此类多重关系重叠无法进行有效标注和关系抽取。

4 实验与结果分析

4.1 数据集及评价标准

金融资讯产生速度快，报道内容复杂多样，且无高质量金融相关开源中文语料库，为保证采集文本的质量以及覆盖率，本文数据来源为东方财富资讯搜索，将主板、中小板以及创业板上市公司名单依次进行搜索爬取，每篇相关资讯的核心内容通常在概述中进行摘要性总结，为避免不存在实体关系对的无关句子增加数据清洗以及人工标注判断的难度，本文选取资讯的概述进行爬取。

东方财富个股资讯信息存在大量的无关资讯，如行情通报类的盘口异动、融资融券变动等，如单一陈述既定事实的强势封涨停板、发布了业绩预告等，这一类资讯占据了大量的篇幅，且不存在可供抽取的实体关系，因此，为最大限度的保证数据集的质量，对数据进行深度清洗后，由3名具有领域知识的人员进行人工筛选以及标注校对工作，共得到标注句1289句，每句平均约有85个字，随机选取其中20%作为测试集，并在训练集中随机选取10%为验证集用作超参数的调整。人工标注后的金融资讯语料如表3所示。

表3 金融资讯语料

	训练集	测试集	共计
标签	86719	22955	109674
普通关系标签	40081	2736	42817
重叠关系标签	1053	268	1321

本文使用的评价指标有查准率(Precision)、召回率(Recall)和F1值(F1-score)对预测结果进行评价，见公式(13)。

$$P=\frac{T_P}{T_P+F_P}\times 100\%$$

$$R=\frac{T_P}{T_P+F_n}\times 100\%$$

$$F1=\frac{2PR}{P+R}\times 100\% \quad (13)$$

式中 T_P 为模型识别正确的关系数，F_P 为模型识别到的不相关关系数，F_n 为相关关系但是没有识别出的个数。对于关系抽取而言，只有关系对中每个实体的实体边界、关系类别和实体角色标签均识别正确时方可获得正确抽取结果。

4.2 实验过程

为证明本实体关系联合抽取模型的有效性，使用本文提出的序列标注模式与实体关系三元组匹配规则和BERT-BiGRU-CRF序列标注模型进行实验，再分别使用BiGRU-CRF与BERT-BiLSTM-CRF序列标注模型进行对比实验。其中

BiGRU-CRF使用随机初始化的embedding矩阵将句中的每个字由one-hot向量映射为低维稠密的字向量作为输入进行训练，用以比对BERT预训练语言模型在本文模型中的效果。其中BERT-BiLSTM-CRF用以比对BiLSTM与BiGRU两种不同的循环神经网络变体的效果。

Google提供的BERT-Base, Chinese预训练语言模型，共12层，隐层具有768维，采用12头模式，参数共有110M个。本文的BERT-BiGRU-CRF序列标注模型训练在Epoch为18时，验证集数据表现不再有明显的提升，所选取的部分主要超参数如表4所示，其中dropout_rate设置为0.3以减少语料偏小情况下的过拟合情况，clip值设置为0.5以防止出现梯度爆炸情况。

表4 BERT-BiGRU-CRF序列标注模型超参数

参数	取值
learning_ rate	5e-5
dimension of BiGRU	256
batch_size	16
clip	0.5
dropout_rate	0.3
max_seq_length	128

4.3 实验结果与分析

使用测试集对本文提出的基于BERT的金融领域实体关系联合抽取模型进行试验，序列标注结果以及实体关系三元组抽取结果如表5所示，“【 】”代表实体，普通关系实体的下标为实体角色标签和实体关系标签，重叠关系实体的下标为重叠关系标签。

按照本文标注模式进行不同序列标注模型对比实验，单个实体、所有实体关系三元组以及包含重叠关系实体的三元组的识别准确率，召回率，F1值如表6所示，其中关系三元组的识别需要关系对中两个实体的标签序列都标注正确方可获得正确的结果。

通过比对我们可以看到单个实体的识别，以及关系抽取上本文采用的模型均取得了较优的效果。其中各模型关系抽取的结果和实体识别的结果相差不大，出现这种情况的原因在于大部分情况下的识别结果为包含关系的两个实体均识别对或均识别错。

对比本文模型与BiGRU-CRF模型，对于实体识别结果，F1值提升了17.0%，在整体关系抽取和重叠关系抽取上，较对照模型F1值分别提升了23.7%和17.3%，说明引入BERT相较于使用常用字向量表示方法更具有优势；对比本文模型与BERT-BiLSTM-CRF，可知GRU虽然结构更为简单，参数更少，但在同条件下可取得不弱于LSTM的效果；本文的标注方案及配对规则使得模型可对重叠关系进行识别抽取，F1值达到了0.543。重叠关系抽取的准确率召回率相较整体关系抽取均偏低，主要在于部分重叠关系存在歧义，存在一句话同时有同一实体的多次表述的情况，使模型较难识别。错误预测如表7所示。

由表7可知，本文使用的序列标注模型将例句2中重复出现的同一实体“融创中国”均标注为重叠关系实体，但是按照本文的抽取配对规则，可以正确抽取出知识三元组。这种现象在模型预测结果中较为罕见，主要原因是实际语料中重叠关系实体出现的位置灵活多变，且句中三处重复的实体均符合本文对重叠关系实体的定义。因此，针对复杂重叠关系，本文提出的联合抽取模型依然存在着改进空间。

5 结 语

本文针对金融领域的复杂资讯文本构造了高质量语料，根据金融文本关系类型及其存在大量的复杂重叠关系的特点，提出了对应的标注模式及实体关系三元组抽取配对规则用以改进对重叠关系的识别，将实体关系抽取转化为序列标注问题；在此基础上使用BERT-BiGRU-CRF序列标注模型进行序列标注以增强字的语义表征和对序列上下文的理解；实现了基于BERT的金融领域实体关系联合抽取模型。实

表5 实体关系联合抽取模型结果举例

例句	标注结果	抽取结果
例1	3月14日，[旭辉控股集团有限公司]$_{重叠实体}$在香港及上海举行2018年度业绩发布会，总裁[林峰]$_{实体1,角色}$携执行副总裁[陈东彪]$_{实体1,角色}$、财务中心总经理[潘皓琦]$_{实体1,角色}$出席上海现场。	(林峰,角色,旭辉控股集团有限公司) (陈东彪,角色,旭辉控股集团有限公司) (潘皓琦,角色,旭辉控股集团有限公司)
例2	[天龙集团]$_{实体1,合作}$(300063)12月18日晚间公告，公司拟与自然人[金华]$_{实体2,合作}$合资设立控股子公司以布局新媒体领域新型业务。	(天龙集团,合作,金华)
例3	[万兴科技]$_{实体1,买卖}$(300624)今日公告拟1.28亿元收购[亿图软件]$_{实体2,买卖}$51%的股权，后者立足于绘图软件领域。	(万兴科技,买卖,亿图软件)

表6 不同序列标注模型实验结果比较

模型	Entity			RE(All)			RE(Overlap)		
	P	R	F1	P	R	F1	P	R	F1
BiGRU+CRF	0.596	0.617	0.606	0.559	0.464	0.507	0.611	0.373	0.463
BERT+BiLSTM+CRF	0.668	0.742	0.703	0.621	0.630	0.626	0.602	0.475	0.531
BERT+BiGRU+CRF	0.670	0.753	0.709	0.620	0.633	0.627	0.583	0.508	0.543

表 7 实体关系联合抽取模型预测错误示例

	正确结果	预测结果
例1标注结果	[富森美]实体1,买卖(002818)晚间公告称,为进一步整合行业资源,拓展精装房和拎包入住业务,公司与[四川建南建筑装饰有限公司]重叠实体(以下简称"四川建南")及其现有股东[廖华]实体1,角色和[廖杰]实体1,角色签署增资协议,公司拟以自有现金向四川建南增资2122.48万元。	[富森美]重叠实体(002818)晚间公告称,为进一步整合行业资源,拓展精装房和拎包入住业务,公司与[四川建南建筑装饰有限公司]实体2,买卖(以下简称"四川建南")及其现有股东[廖华]实体1,角色和[廖杰]实体1,角色签署增资协议,公司拟以自有现金向四川建南增资2122.48万元。
例1抽取结果	(富森美,买卖,四川建南建筑装饰有限公司) (廖华,角色,四川建南建筑装饰有限公司) (廖杰,角色,四川建南建筑装饰有限公司)	(富森美,买卖,四川建南建筑装饰有限公司) (廖华,角色,富森美) (廖杰,角色,富森美)
例2标注结果	12月5日,[融创中国]重叠实体康养品牌发布盛典在山东青岛举行,融创中国董事会主席[孙宏斌]实体1,角色、融创中国执行董事兼行政总裁[汪孟德]实体1,角色等人出席。	12月5日,[融创中国]重叠实体康养品牌发布盛典在山东青岛举行,[融创中国]重叠实体董事会主席[孙宏斌]实体1,角色、[融创中国]重叠实体执行董事兼行政总裁[汪孟德]实体1,角色等人出席。
例2抽取结果	(孙宏斌,角色,融创中国) (汪孟德,角色,融创中国)	(孙宏斌,角色,融创中国) (汪孟德,角色,融创中国)

验结果初步验证了中文语境下本文所提出的联合抽取模型对金融领域实体关系抽取的有效性。

但本文提出的联合抽取模型仍然不能对更为复杂的关系进行有效抽取,实际场景中人工构造的高质量语料很难达到足够的规模会影响模型效果,预定义的关系也较难满足当前呈爆发状态信息增长的信息抽取任务要求。因此,如何在小规模高质量语料的情况下提升模型的学习能力,提升对复杂关系的识别能力,并能够从句子内的实体关系抽取,到句子与句子或者段与段之间这样的复杂语境情况下进行有效的实体关系抽取,还有待进一步的研究探索。

参考文献

1 DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J].arXiv preprint arXiv: 181004805, 2018,

2 YADAV V, BETHARD S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models[EB/OL].https://ui.adsabs.harvard.edu/abs/2019arXiv191011470Y,2019-06-05.

3 刘 浏, 王东波. 命名实体识别研究综述[J].情报学报, 2018, 37(3): 329-40.

4 CUCERZAN S, YAROWSKY D. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence[C]//proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, F, 1999.

5 ZHOU G, SU J. Named entity recognition using an HMM-based chunk tagger[M]. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania; Association for Computational Linguistics,2002: 473 - 80.

6 DUAN H, ZHENG Y. A study on features of the CRFs-based Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence, 2011, 3(2): 287-294.

7 LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J].arXiv preprint arXiv:160301360, 2016,

8 STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J].arXiv preprint arXiv:170202098, 2017.

9 李枫林, 柯 佳. 基于深度学习框架的实体关系抽取研究进展[J].情报科学, 2018, 36(3): 169-176.

10 车万翔, 刘 挺, 李 生. 实体关系自动抽取[J].中文信息学报, 2005,(2): 1-6.

11 ZHANG M, ZHANG J, SU J, et al. A composite kernel to extract relations between entities with both flat and structured features[M]. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia:Association for Computational Linguistics,2006: 825 - 832.

12 LIU C, SUN W, CHAO W, et al. Convolution neural network for relation extraction[C]//proceedings of the International Conference on Advanced Data Mining and Applications, F,Springer,2013.

13 ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//proceedings of the Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), F, 2016.

14 CAI R, ZHANG X, WANG H. Bidirectional recurrent convolutional neural network for relation classification [C]//proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), F, 2016.

15 ZHENG S, WANG F, BAO H, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme

[J]. ACI,2017,(6):438-450.
16 YU X, LAM W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach [C]//proceedings of the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, F, Association for Computational Linguistics,2010.
17 LI Q, JI H. Incremental joint extraction of entity mentions and relations[C]//proceedings of the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), F, 2014.
18 曹明宇, 杨志豪, 罗　凌. 基于神经网络的药物实体与关系联合抽取[J].计算机研究与发展,2019,56(7):1432-1440.
19 MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//proceedings of the Advances in neural information processing systems, F, 2013.
20 PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), F, 2014.
21 王树伟. 面向金融文本的实体识别与关系抽取研究[D].哈尔滨:哈尔滨工业大学,2014.
22 孔　兵. 中文文本实体关系抽取方法研究 [D].哈尔滨:哈尔滨工业大学, 2016.
23 RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL].URL https://s3-us-west-2 amazonaws com/openai-assets/researchcovers/languageunsupervised/language understanding paper pdf, 2018-03-02,
24 PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J].arXiv preprint arXiv: 180205365, 2018,
25 杨　飘,董文永.基于 BERT 嵌入的中文命名实体识别方法[EB/OL].https://doi.org/10.19678/j.issn.1000-3428.005054272,2020-04-01.
26 CHO K, VAN MERRIëNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J].arXiv preprint arXiv:14061078, 2014,
27 HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J].Neural computation, 1997,9(8): 1735-1780.
28 GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J].Neural networks,2005,18(5-6): 602-610.

（责任编辑：赵红颖）