期货领域知识图谱构建*

李雯昕[1,2]，眷红英[1]，关同峰[1,3]，韩英杰[1]
1.郑州大学计算机与人工智能学院，河南郑州
2.中移在线服务有限公司，河南郑州　3.中原银行，河南郑州
wency.li@foxmail.com; iehyzan@zzu.edu.cn;
guantf.gtf@foxmail.com; ieyjhan@zzu.edu.cn

摘要

期货领域是数据最丰富的领域之一，本文以商品期货的研究报告为数据来源构建了期货领域知识图谱（Commodity Futures Knowledge Graph，CFKG）。以期货产品为核心，确立了概念分类体系及关系描述体系，形成图谱的概念层；在MHS-BIA与GPN模型的基础上，通过领域专家指导对242万字的研报文本进行标注与校对，形成了CFKG数据层，并设计了可视化查询系统。所构建的CFKG包含17,003个农产品期货关系三元组、13,703种非农产品期货关系三元组，为期货领域文本分析、舆情监控和推理决策等应用提供知识支持。

关键词： 知识图谱；命名实体识别；实体关系抽取；期货文本

Construction of Knowledge Graph in Futures Field

Wenxin Li [1,2], Hongying Zan[1], Tongfeng Guan[1,3], Yingjie Han[1]
1.School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan, China
2.China Mobile Online Services, Zhengzhou, Henan, China　3.Zhongyuan Bank, Zhengzhou, Henan, China
wency.li@foxmail.com; iehyzan@zzu.edu.cn;
guantf.gtf@foxmail.com; ieyjhan@zzu.edu.cn

Abstract

The futures field is one of the fields with the most abundant data.This paper used the Research Report of commodity futures as the data source to construct the Commodity Futures Knowledge Graph (CFKG). Taking futures products as the core, the concept classification system and the relationship description system were established, thus the CFKG conceptual layer was also built.Based on MHS-BIA and GPN model, more than 2.42 million words of research reports were annotated and proofread under the guidance of field experts, forming the CFKG data layer,and designed a visual query system of the knowledge graph. The futures domain knowledge graph constructed in this thesis containing 17,003 agricultural product relation triples and 13,703 non-agricultural product relation triples, which provided structured knowledge support for text analysis, public opinion monitoring, reasoning and decision-making in the futures field.

Keywords: Knowledge Graph , Named Entity Recognition , Relation Extraction , Futures Text

期货市场智能化舆情分析研究项目（20200464A）

第二十一届中国计算语言学大会论文集，第246页-第256页，南昌，中国，2022年10月14日至16日。

1 引言

期货行业是个数据驱动的行业，随着期货市场逐渐扩大，期货领域相关数据和类型不断增多，从大数据中挖掘有价值的信息并应用于期货领域是必然趋势。该领域数据具有数据信息密集、数据量庞大、数据种类多样等特征，传统业务对行情、咨讯数据已经形成了高度依赖。但是，也存在数据利用效率低，对现有的数据资源价值的挖掘、分析和利用能力较弱的问题。因此，使用实体识别和关系抽取技术将实体、属性、关系等从非结构化、半结构化数据中抽取出来并建立关联，从内容分散、多元异构的期货文本中挖掘重要信息形成期货知识图谱，有助于金融从业者高效获取信息，为期货领域文本分析、舆情监控等关键技术提供坚实的基础，促进期货领域智能化发展。

知识图谱可提供一种更好的组织和理解信息的能力，2012年5月Google正式提出知识图谱的概念。发布基于维基百科(Vrandečić and Krötzsch, 2014)、Freebase(Bollacker et al., 2008)的知识图谱。根据数据源可将知识图谱划分为通用和领域知识图谱，通用知识图谱以百科类网站为数据源，面向公共领域，如跨语言百科知识图谱XLore(Wang et al., 2013)，中文模式知识库Zhishi.me(Niu et al., 2011)等。领域知识图谱依托特定专业领域的数据进行构建，在知识图谱构建过程中依赖领域专家或工程师依据项目的具体应用背景和需求进行图谱规则的制定，追求领域知识的专业性和准确性，如医学领域(奥德玛et al., 2019)、电商领域(Xu et al., 2021)和旅游领域(Kärle et al., 2018)等。将知识图谱应用于期货领域，可链接多数据源，形成商品和用户的知识描述体系，从而让商家和用户更直观地了解、认识、分析用户群体和商品。虽然目前学术界对于该领域知识图谱的研究还不多，就相关领域而言，爱智慧科技有限公司构建产业链图谱(Chen et al., 2021)，图谱包括有色金属和非金属材料等产品的行业结构与产品上下游信息。

期货行业除了对期货涉及的交易品种、合约、期货公司等实体外，还需对期货相关实体更大粒度的结构化标签如宏观经济主题、交易品种类别等进行识别，对期货相关实体细粒度的结构化标签等也需要进行识别。因此，以上相关研究无法满足期货领域知识图谱分析需求。由于期货领域相关企业内部的信息属于企业机密，可供研究的高质量公开数据较少，并且期货数据更新速度快、信息密集，无法及时进行知识更新及挖掘补充。期货市场尚未有成熟的信息检索产品进行大规模投放，相关研究正处于探索阶段，说明研究期货领域知识图谱的构建方法的重要性和迫切性。针对以上问题，本文针对期货领域的特点，针对商品期货展开研究，采用实体识别和关系抽取技术从非结构化期货文本中抽取出实体和关系，构建期货领域知识图谱（Commodity Futures Knowledge Graph，CFKG）。

2 CFKG构建流程

知识图谱从逻辑结构上可划分为概念层和数据层两个部分。概念层对实体概念和关系分类体系进行建模，对数据层进行约束。数据层通过实体关系三元组对概念层各类知识的定义进行表达。

CFKG构建过程如图1所示，首先设计概念层，即制定相应的知识描述体系。根据期货公司发布的期货研报中收集期货领域语料与术语集合，通过示例标注与分析设计实体概念体系和关系分类体系，经领域专家评估后形成知识图谱描述体系。以概念层为基础，获取多来源期货领域文本，采用半自动的方式对实体及实体关系进行标注构建语料库（Corpus for Entity and Relation annotation in Futures domain，CERF）。并展开期货实体关系联合抽取模型的研究，采用自动方式实现知识的自动更新，CERF语料库与模型自动抽取的实体及关系三元组经数据整合后形成CFKG的数据层。

3 CFKG概念层构建

本文根据期货产品的特点将其划分为农产品期货和非农产品期货两种，分别构建概念分类描述体系和关系分类描述体系。CFKG以期货产品作为描述主体，通过建立与产品相关的价格、地名、质量指标等概念之间的联系，形成各类概念之间的网状关系。CFKG中定义了产品、企业、价格、地名、价格指标等20类农产品期货实体与14类非农产品期货实体，对期货产

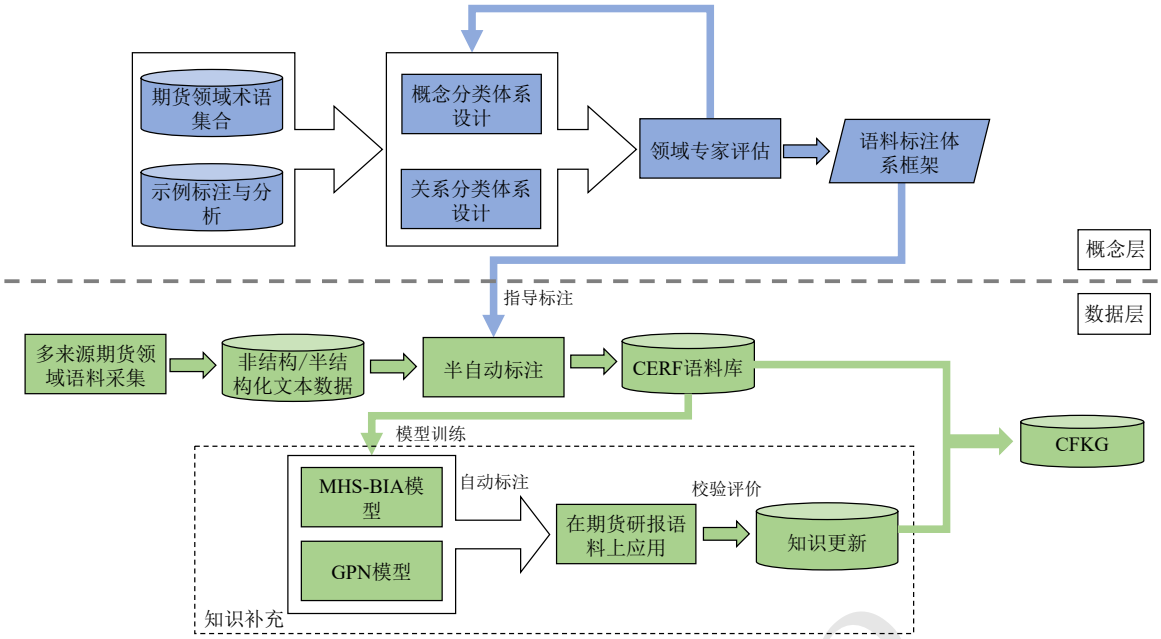

图 1. CFKG构建流程图

品研究报告中各类产品描述进行考察，在概念分类体系基础上进一步设计实体间关系描述体系，以产品作为关系描述中心拓展与其他期货实体间的关系。农产品期货包含9类关系类型，包括42种子关系；非农产品期货包含11类关系类型，包含36种子关系。由于篇幅所限，表1与表2仅展示了CFKG中农产品期货概念分类体系及关系分类体系，其余子关系描述体系不再一一详述。

关系类型	子关系	关系定义
产品-产品	同义	包括别名、简称、英文缩写以及同义词
	替代	在生产、流通或消费环节的可代替品
	下游	指以该产品为原材料生产或加工的产品
	AKO	A kind of：泛指一个实体是另一个实体中的一类

表 1. CFKG部分标注关系定义

4 CFKG数据层构建

CFKG数据层构建主要包括期货领域语料采集、半自动标注及模型自动抽取三部分。

4.1 领域数据采集

期货领域缺少专用的行业规范资源，如专用术语标准、行业指南等可供参考，因此需要从多方面获取期货文本资源，形成期货领域知识语料库。中文期货领域文本包括期货品种简介、期货新闻、期货研报等网络资源，不同来源的数据在文本内容上存在差异，如期货研报是我国期货业内专家针对不同期货品种和行业动态做出的分析和研究；期货品种简介以归纳方式总结各类期货品种的特点；期货新闻侧重于期货关联商品行业动态和发展状况。CFKG以产品为核心，抽取多来源文本中的相关内容，期货领域数据来源如表2所示，其中“半”表示半结构化数据，“非”表示非结构化数据。

通过对比分析数据收集阶段获得的语料，期货研报相比新闻文本数据中包含对不同期货品种的行业动态、产业链信息做出分析和研究，且由专业的期货分析师撰写，可信度高，专业性强，因此选取研报数据作为人工标注的语料。并取期货文本较为丰富的六个期货品种：棉花、

名称	描述	形式	语料规模/篇
研究报告	期货领域行业研究人员针对不同的期货品种和行业动态做出分析和研究;	半/非	634
品种简介	郑商所官网提供的已上市的期货品种简介，包含期货领域知识、术语和规范;	半	195
新闻文本	行业内期货分析师团队编写，提供每日国内外期货行情、期货报价等;	半	1,182

表 2. 期货领域多来源数据采集

苹果、白糖、红枣、玻璃、PTA作为研究对象。从Wind金融终端获取的研报为PDF格式，原始语料经预处理后的待标注语料规模总计174万字，包含74,437个句子。

4.2 半自动标注

半自动标注指根据期货产品的实体和关系分类体系，使用基于词典库的双向最大匹配算法对语料进行预标注，在此基础上进行人工标注和校对。为了提升标注效率，借鉴医学关系标注平台(张坤丽et al., 2020)，结合期货领域知识进行重新配置，形成了面向期货领域的实体和关系标注平台。其中词典库根据表2采集的数据经半自动标注-专家确认形成。预先在标注平台配置期货产品的实体和关系分类体系，在标注关系时仅可选择预定义的实体和关系类型，使用不同颜色表示不同的实体概念，使用连线和类型标签表示实体间关系类型。具体的标注流程如图2所示。 标注一致性用来描述两份标注结果的一致程度，一般使用Kappa值(Carletta,

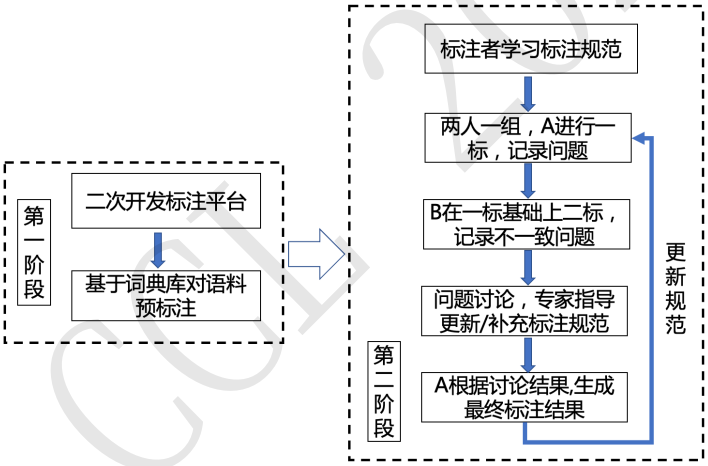

图 2. 半自动标注过程

1996)和F1值(Hripcsak and Rothschild, 2005)进行评价。文献(Artstein and Poesio, 2008)指出，标注一致性高于80%，可认为标注语料是可靠的，统计结果显示CERF语料库中实体和关系标注一致率F1值分别为85.70%和81.80%，均高于80%，说明本文构建的标注语料库可靠。

4.3 CERF语料库构建结果

期货领域文本包含丰富的实体关系三元组，因此单个实体可能同时涉及多个三元组中，进而导致实体出现重叠。实体重叠问题可分为普通型（NOR）、单个实体重叠型（SEO）和实体对重叠型（EPO）。普通型指无实体重叠问题；当单句中存在某个实体与其他多个实体存在关系，属于单个实体重叠型；实体对重叠型指句子中的相同实体对存在多种语义关系。CERF标注语料库规模如表3、表4所示，其中，实体重叠类型总数与语料总数不一致，原因为同一条语料中可能同时出现实体对重叠类型和单个实体重叠型两种类型的三元组。

产品类别	农产品			非农产品		
语料类别	训练集	验证集	测试集	训练集	验证集	测试集
语料数目	9,652	2,758	1,388	2,765	790	401

表 3. CERF实体语料库规模

产品类别	农产品			非农产品		
语料类别	训练集	验证集	测试集	训练集	验证集	测试集
子关系数目	42	42	42	36	36	36
语料数目	3,216	402	402	2,526	316	316
三元组数目	8,976	1,110	1,223	8,723	1,014	1,095
#单个句子中三元组数目						
1	1,173	114	136	758	105	97
2	741	104	94	488	63	65
3	488	54	51	373	48	44
4	292	31	35	282	32	25
≥ 5	522	69	86	625	68	85
#实体重叠类型						
NOR	1,398	171	164	873	123	115
SEO	1,818	231	238	1,653	193	201
EPO	666	33	64	167	11	9

表 4. CERF关系语料库规模

4.4 实体及关系抽取算法研究

在CERF标注语料库基础上，展开期货领域实体识别和关系抽取模型研究，将模型应用于期货研报文本进行实体识别及关系抽取，将数据整合后得到结构化三元组作为CFKG数据层的补充。（1）MHS-BIA模型

针对三元组重叠问题，本节在Bekoulis et al. (2018a)和Zhang et al. (2016)等研究基础上，提出基于Biaffine注意力的多头选择模型（Multi-Head Selection based on BIAffine attention，MHS-BIA）的改进算法，将信息抽取问题定义为一个多头选择问题(Bekoulis et al., 2018b)。模型能够同时识别实体，包括实体类型和实体边界，以及实体对之间所有可能存在的关系，使用Sigmoid损失获得实体间多个语义关系，以此能够独立预测不互斥的类。具体做法为：将模型识别的实体中最后一个字符称为实体“头”，如产品实体“烟台苹果”的尾字符“果”。模型认为该实体与本句中其他任意实体均可能存在语义关系。依次判别实体尾字符x_i其他实体尾字符y_l之间是否存在语义关系$\hat{c}_l$,若存在语义关系则将判断结果记作元组$(\hat{y}_l, \hat{c}_l)$。对于没有语义关系的元组，引入“N”标记无关系。模型框架如图3所示，包括实体识别模块和关系分类模块。

实体识别模块中的多头选择网络思想是将输入序列$W = \{w_1, w_2, ..., w_n\}$中每个字符$w_j, j \in \{1, 2, ..., n\}$组合，进而判断这两个字符是不是某个实体的头尾字符，并且将其归属于预定义实体类型e_k。通过将两个编码特征矩阵拼接，再通过线性变换层。计算w_i和w_j之间组成实体且实体关系为e_k的分数公式如（1）和（2）所示:

$$d = \dot{c}_i + \dot{c}_j \tag{1}$$

$$s^{(e)}(i, j, k) = \delta(Wd + b) \tag{2}$$

其中，d为两个编码特征矩阵拼接后的结果，c_i 和c_j表示BERT编码层的输出序列经过一层线性层变换和激活函数后得到的编码表示，$\dot{c}_i$ 和$\dot{c}_j$表示复制n份后得到的编码特征表示。W和b分别表示参数权重和偏差，δ为Relu激活函数。

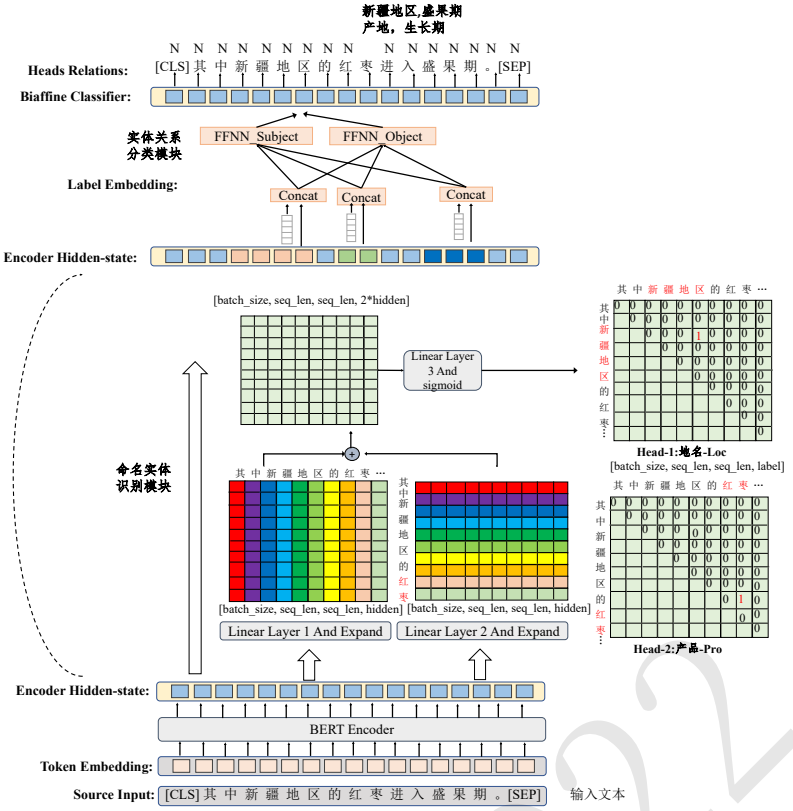

图 3. MHS-BIA模型

实体识别的交叉熵损失函数，与字符w_i和w_j之间存在关系e_k的可能性计算公式分别如（3）和（4）所示:

$$L_{ent} = \sum_{i=0}^{n} \sum_{j=0}^{m} -\log Pr\left(head = y_{i,j}, ent\_type = e_{i,j} \mid w_i\right) \tag{3}$$

$$Pr\left(head = w_j, ent\_type = e_k \mid w_i\right) = \sigma\left(s^{(e)}(i,j,k)\right) \tag{4}$$

$\sigma(\cdot)$为sigmoid函数，n为输入序列长度，m包括字符w_i的实体数量，$y_i \subseteq W$为除了w_i之外的其他字符，e_k表示实体关系。若两个字符是某实体类型下实体的头尾字符则为1，反之为0。

关系分类模块和实体识别模块共享同一个特征抽取层，将关系模块学习的特征与实体软标签进行向量拼接，结合两个子任务之间的联系。模型的任务为，对于给定的输入文本序列W和预定义好的关系集合$\mathcal{R}$，正确预测实体尾字符w_i,$i \in \{0,1,...,n\}$与其他实体尾字符所存在的语义关系$\hat{r}_i \subseteq \mathcal{R}$ 计算两个实体尾字符之间存在关系r_k的得分公式如（5）至（7）所示:

$$g_i = \frac{\sum soft\max\left(s_i\right) \cdot \mathrm{M}}{N} \tag{5}$$

$$z_i = \left[h_i; g_i\right], i = 0, \ldots, n \tag{6}$$

$$s\left(z_j, z_i, r_k\right) = V \cdot \mathrm{f}\left(U z_j + W z_i + b\right) \tag{7}$$

式（5）表示学习实体标签向量表示g_i，其中s_i代表输入序列第i个字符的状态分数向量，N代表预定义的实体标签种类数目，M为标签向量矩阵。式（6）表示向量拼接过程，z_i为特征抽取层的输出表示h_i与实体标签向量表示g_i两者拼接后的向量表示。式（7）表示计算关系系数的过程，其中$f(\cdot)$为relu激活函数，$V, b \in \mathbb{R}^l$,b代表向量维度，l代表当前层隐藏单元数。$U, W \in \mathbb{R}^{l \times (2d+b)}$，d为编码层隐藏单元数量。

传统浅层双线性分类器使用特征提取层的输入直接参与下一步计算，缺点在于模型任意时刻的输出均包含其他时刻的信息，本文在关系分类模块引入Biaffine注意力机制取代浅层双线性分类器，计算实体头尾字符的向量表示。深层Biaffine注意力机制模型将抽取的特征传入前馈神经网络，增加了本身的偏差项。Biaffine计算如公式（8）-（10）所示:

$$z_i' = FFNN_{Subject}\left(z_i\right) \tag{8}$$

$$z_j' = FFNN_{Object}\left(z_j\right) \tag{9}$$

$$s_m\left(z_i', z_j'\right) = z_i' U_m z_j' + W_m\left(z_i' \oplus z_j'\right) + b_m \tag{10}$$

其中z_i和z_i为式（6）中特征抽取的输出与实体软标签向量的拼接层，$FFNN_{Subject}$和$FFNN_{Object}$为前馈神经网络，z_i'和z_j'表示降维后的结果。实体对在所有关系类型上的得分计算如式（10）所示，b_m为偏差，$U_m \in \mathbb{R}^{d \times c \times d}$，$W_m \in \mathbb{R}^{(2d \times c)}$，d代表前馈神经网络隐藏单元数，c为关系类型数量。预测实体尾字符之间所有关系类型的概率值如式（11）所示，交叉熵损失函数如式（12）所示。其中n为序列长度，m为实体尾字符所涉及三元组数目，$r_i \in \mathcal{R}$为实体对的语义关系。

$$Pr\left(head = w_j \mid w_i\right) = softmax\left(s_m\left(z_i', z_j'\right)\right) \tag{11}$$

$$L_{rel} = \sum_{i=0}^{n}\sum_{j=0}^{m} -\log Pr\left(head = y_{i,j}, relation = r_{i,j} \mid w_i\right) \tag{12}$$

（2）GPN模型

基于全局指针网络的实体关系联合抽取模型（Joint extraction model based on Global Pointer Networks, GPN）在TPLinker模型(Wang et al., 2020)基础上，引入全局指针思想联合解码，将标注抽取框架设定为字符对链接问题，识别实体时将实体首位字符视作一个整体进行判别，最后使用条件层归一化的方式取代句子编码和主实体信息简单相加的方式，解决暴露偏差问题。

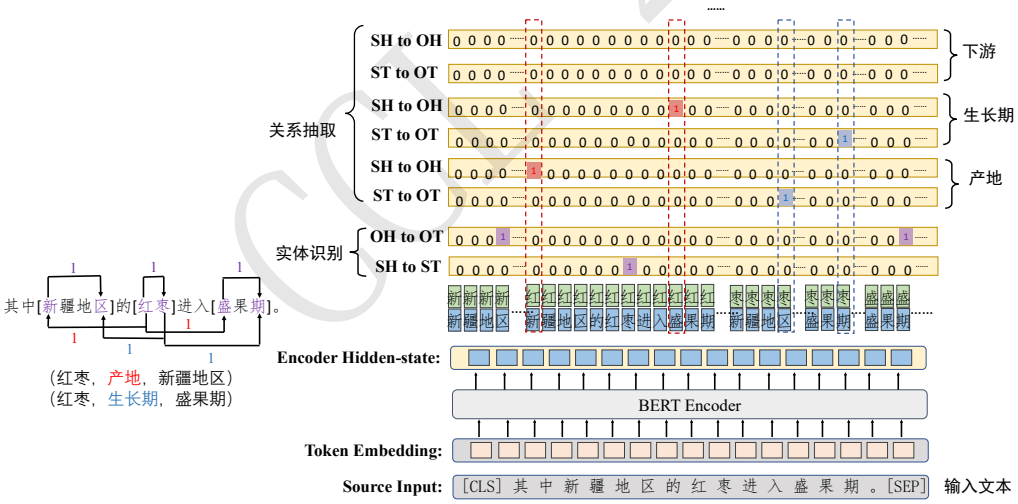

图 4. GPN模型

关系抽取的结果通常由三元组（Subject, Predicate, Object）表示，但模型在抽取过程实际为“五元组”（S_h,S_t,P,O_h,O_t）的抽取，其中S_h,S_t分别表示主实体的首尾位置，O_h,O_t分别表示客实体首尾位置。GPN模型使用单阶段实体关系联合抽取的标注方案，通过对主、客实体的头字符和尾字符进行标记，将实体识别和关系抽取描述为token对链接问题。模型标注框架如图4所示，分为实体识别和关系抽取两个模块，通过链接三种类型的矩阵实现标注过程。

编码阶段：1.主、客实体识别：SH to ST（Subject Head to Subject tail）表示识别主实体首尾字符；OH to OT（Object Head to Object tail）表示识别客实体的首尾字符；2.主、客实体头token识别：SH to OH，表示识别主、客实体首字符间的关系，如三元组（红枣，产地，新疆地区）：（红，新）=1；3.主、客实体尾token识别：ST to OT，表示识别主、客实体尾字符间的关系，如三元组（红枣，产地，新疆地区）：（枣，区）=1。

给定长度为n的输入序列$W = (w_1, w_2, \ldots, w_n)$，通过BERT等预训练模型将字符序列编码为向量序列h_N表示，其中每个字符w_i,$i \in \{1, 2, ..., n\}$映射到低维的语义向量$x_i = h_N[i]$,$i \in \{1, 2, \ldots, n\}$。GPN模型采用缩放点积型注意力机制，其字符对（$w_i, w_j$）的分类特征$x_(i,j)$计算公式如式（13）所示，其中$x_i, x_j \in {}_k$为字符的语义向量表示，$d_k$为向量维度，$x_{i,j}^{(\cdot)}$为不同类型的字符矩阵中值的分数表示。

$$x_{i,j}^{(\cdot)} = \frac{x_i^T x_j}{\sqrt{d_k}} \tag{13}$$

GPN模型为主实体的首尾（SH-ST）、客实体的首尾（OH-OT）、主实体和客实体的首字符对间的关系（SH-OH）、主实体和客实体的尾字符对间的关系（ST-OT）标注设计统一的框架。给定一个字符对（w_i, w_j）的特征表示$x_{i,j}^{(\cdot)}$，其链接标签计算公式如（14）和（15），其中$P\left(y_{i,j} = l\right)$表示（$w_i, w_j$）之间链接被识别为l的概率。

$$P\left(y_{i,j}\right) = Softmax\left(W \cdot x_{i,j}^{(\cdot)} + b\right) \tag{14}$$

$$link\left(w_i, w_j\right) = P_{argmax}\left(y_{i,j} = l\right) \tag{15}$$

解码阶段：1.解码SH-ST、OH-OT可得到句子中所有的实体，将主实体首字符索引作为关键字，整个实体作为值存储在字典D中；2.对于关系分类，解码SH-OH后，得到主实体和客实体的首字符token对，并在字典D中关联首字符索引链接的实体值。解码ST-OT后，得到主、客实体尾字符token对存储于集合T中；3.对第2步获取的主、客实体头字符token对链接到实体对，遍历集合T查询是否存在其尾字符token，若存在则输出实体关系三元组。

（3）实验结果

使用MHS-BIA模型与GPN模型在CERF语料库上进行实体关系抽取实验，并使用BiLSTM-CRF模型(Huang et al., 2015)和CASREL模型(Wei et al., 2019)作为对比模型。设置词向量维度为300，位置向量维度为300，Dropout 为0.5，GPN实体关系抽取结果见表6和表7。针对CERF语料库中部分实体类别标注错误，非农产品数据标记样本少等问题，对实体语料库训练集做了数据增强处理(Wei and Zou, 2019)，包括根据频次修正实体类型，简单数据增加等。

模型	农产品			非农产品		
	Micro-P (%)	Micro-R (%)	Micro-F1 (%)	Micro-P (%)	Micro-R (%)	Micro-F1 (%)
BiLSTM-CRF	75.00	76.67	75.82	70.70	76.98	73.71
CASREL	79.54	72.84	76.04	75.29	74.59	74.93
MHS-BIA	76.92	76.65	76.79	71.43	76.56	73.91
MHS-BIA-EDA	77.24	73.85	76.52	74.67	76.93	75.78
GPN	79.44	74.59	**76.94**	77.35	72.16	74.66
GPN-EDA	77.55	74.76	76.13	79.03	73.72	**76.28**

表 5. 模型命名实体识别实验结果

命名实体识别实验结果表明数据增强的方法更适用于训练数据少的任务，比如非农产品语料库中，数据增强的方法获得了1.62%的提升。而在农产品语料中，数据增强的方法并未获得提升。可能的原因为本章采用增加简单句的方法扩充训练数据，农产品的训练数据较非农产品更充足，增加数据样本的方法并不能提升效果。因此过多的数据增强对模型提升有限，该方法对训练数据越小的任务提升效果越明显。

实体关系抽取实验结果表明GPN模型结果优于其他模型，而MHS-BIA模型相比GPN的不足，可能原因在于实体和关系类别数目的限制：当实体或关系类别数目过多时，产生信号稀疏

模型	农产品			非农产品		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
CASREL	53.49	52.26	52.87	54.38	49.84	52.01
MHS-BIA	57.43	52.31	54.75	56.00	50.08	52.87
GPN	59.92	52.79	**56.13**	60.64	48.13	**53.67**

表 6. 模型实体关系抽取实验结果

问题，导致模型监督信号减弱，训练难度增大；此外，实体和关系任务抽取时共用一个编码器，使得关系分类判别依然依赖实体识别结果，存在暴露偏差，最终导致误差积累。本文未在关系抽取实验中使用数据增强技术，因为相较于实体识别，关系抽取语料的句子包含特定的关系三元组，多数语料文本长度更长。而自动构造的相似句长度较短并且生成的三元组质量不高，无法满足关系抽取任务，后续可考虑通过远程监督扩充关系抽取的语料。

5 构建结果及展示

CFKG的数据层由半自动标注构建的CERF语料库和模型自动标注结果两部分组成。本文抽取年份为2020.10-2022.1的六种期货产品的研报作为数据补充，采用实验效果较好的GPN模型进行实体关系联合抽取，经数据整合后得到结构化的实体关系三元组，作为CFKG数据层的补充。图谱中实体关系数量如表7、表8所示，共包含17,003个农产品关系三元组、13,703种非农产品关系三元组。

关系类型	半自动标注	自动抽取	合计
产品-属性	2,680	873	3,553
产品-地名	1,995	1,081	3,076
产品属性-变化	2,195	1,744	3,939
价格-因素	1,033	275	1,308
期货术语-属性	963	589	1,552
产品-价格	787	226	1,013
产品-产品	661	102	763
产品-因素	595	731	1,326
产品-其他	400	73	473

表 7. CFKG中农产品实体关系三元组统计

关系类型	半自动标注	自动抽取	合计
产品-属性	3,520	1,205	4,725
产业链条件-因素	3,358	777	4,135
价格-因素	805	180	985
期货术语-属性	681	199	880
行业-因素	649	60	709
生产设施-产业链条件	564	92	656
产品-地名	478	96	574
产品-企业	322	223	545
产品-产品	219	28	247
期货品种-期货术语	141	11	152
其他指标-属性	95	0	95

表 8. CFKG中非农产品实体关系三元组统计

为了直观反应CFKG中实体之间的关系，设计了期货领域知识图谱可视化查询系统。系统通过问句解析模块与知识图谱检索模块，实现对CFKG中节点进行查询和检索，展示界面如

图5所示。

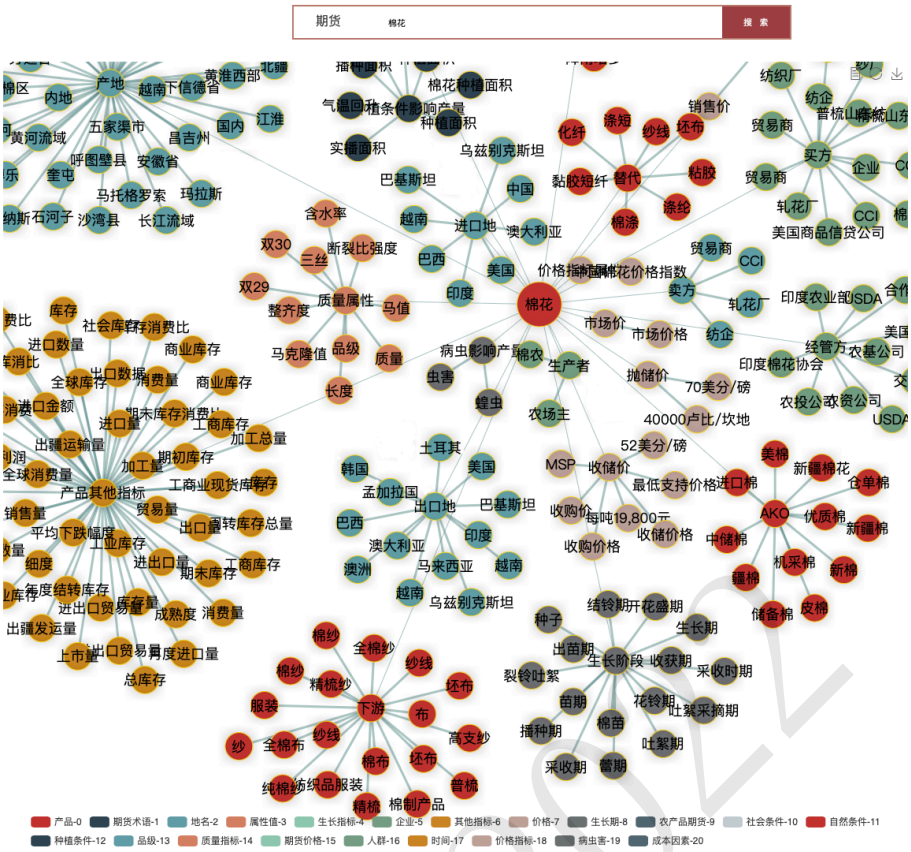

图 5. 棉花产品知识图谱展示

6 结语

本文描述了期货领域知识图谱CFKG的构建过程。首先在概念层整合多来源的期货文本，经领域专家指导下设计了知识图谱描述体系。在数据层采用半自动标注的方法构建了期货领域实体及关系标注语料库，使用自主构建的GPN模型进行自动标注，实现CFKG的知识更新。最后设计了期货领域知识图谱可视化查询系统对图谱进行可视化展示。未来的研究工作将在提升CFKG数据质量的同时，展开基于小规模语料的模型研究，或使用现有的信息抽取模型抽取新的语料作为训练语料补充，促使模型抽取效果迭代提升。其次本文构建的知识图谱属于静态的知识表示，但期货领域相比于传统领域信息更新快，故对知识更新速度具有较高的需求，未来工作可针对期货领域事理图谱展开研究，以及在静态知识图谱中融入动态知识实现图谱的动态更新。

参考文献

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100–112.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004*.

Huajun Chen, Ning Hu, Guilin Qi, Haofen Wang, Zhen Bi, Jie Li, and Fan Yang. 2021. Openkg chain: A blockchain infrastructure for open knowledge graphs. *Data Intelligence*, 3(2):205–227.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Elias Kärle, Umutcan Şimşek, Oleksandra Panasiuk, and Dieter Fensel. 2018. Building an ecosystem for the tyrolean tourism knowledge graph. In *International Conference on Web Engineering*, pages 260–267. Springer.

Xing Niu, Xinruo Sun, Haofen Wang, Shu Rong, Guilin Qi, and Yong Yu. 2011. Zhishi. me-weaving chinese linking open data. In *International Semantic Web Conference*, pages 205–220. Springer.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *International semantic web conference (Posters & Demos)*, volume 1035, pages 121–124.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A novel cascade binary tagging framework for relational triple extraction. *arXiv preprint arXiv:1909.03227*.

Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. 2021. Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4808–4812.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2016. Dependency parsing as head selection. *arXiv preprint arXiv:1606.01280*.

奥德玛, 杨云飞, 穗志方, 代达劢, 常宝宝, 李素建, and 昝红英. 2019. 中文医学知识图谱cmekg 构建初探. 中文信息学报, 33(10):1–7.

张坤丽, 赵旭, 关同峰, 尚柏羽, 李羽蒙, and 昝红英. 2020. 面向医疗文本的实体及关系标注平台的构建及应用. 中文信息学报, 34(6):36–44.