软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(7):3377−3391 [doi: 10.13328/j.cnki.jos.006931]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向合同信息抽取的动态多任务学习方法*

王浩畅[1], 郑冠彧[1], 赵铁军[2]

[1](东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

[2](哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

通信作者: 王浩畅, E-mail: kinghaosing@gmail.com

摘　要: 对于合同文本中要素和条款两类信息的准确提取, 可以有效提升合同的审查效率, 为贸易各方提供便利化服务. 然而当前的合同信息抽取方法一般训练单任务模型对要素和条款分别进行抽取, 并没有深挖合同文本的特征, 忽略了不同任务间的关联性. 因此, 采用深度神经网络结构对要素抽取和条款抽取两个任务间的相关性进行研究, 并提出多任务学习方法. 所提方法首先将上述两种任务进行融合, 构建一种应用于合同信息抽取的基本多任务学习模型; 然后对其进行优化, 利用 Attention 机制进一步挖掘其相关性, 形成基于 Attention 机制的动态多任务学习模型; 最后针对篇章级合同文本中复杂的语义环境, 在前两者的基础上提出一种融合词汇知识的动态多任务学习模型. 实验结果表明, 所提方法可以充分捕捉任务间的共享特征, 不仅取得了比单任务模型更好的信息抽取结果, 而且能够有效解决合同文本中要素与条款间实体嵌套的问题, 实现合同要素与条款的信息联合抽取. 此外, 为了验证该方法的鲁棒性, 在多个领域的公开数据集上进行实验, 结果表明该方法的效果均优于基线方法.

关键词: 多任务学习; 合同文本; 信息联合抽取; 注意力机制; 实体嵌套

中图法分类号: TP18

中文引用格式: 王浩畅, 郑冠彧, 赵铁军. 面向合同信息抽取的动态多任务学习方法. 软件学报, 2024, 35(7): 3377–3391. http://www.jos.org.cn/1000-9825/6931.htm

英文引用格式: Wang HC, Zheng GY, Zhao TJ. Dynamic Multitask Learning Approach for Contract Information Extraction. Ruan Jian Xue Bao/Journal of Software, 2024, 35(7): 3377–3391 (in Chinese). http://www.jos.org.cn/1000-9825/6931.htm

Dynamic Multitask Learning Approach for Contract Information Extraction

WANG Hao-Chang[1], ZHENG Guan-Yu[1], ZHAO Tie-Jun[2]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Accurately extracting two types of information including elements and clauses in contract texts can effectively improve the contract review efficiency and provide facilitation services for all trading parties. However, current contract information extraction methods generally train single-task models to extract elements and clauses separately, whereas they do not dig deep into the characteristics of contract texts, ignoring the relevance among different tasks. Therefore, this study employs a deep neural network structure to study the correlation between the two tasks of element extraction and clause extraction and proposes a multitask learning method. Firstly, the primary multitask learning model is built for contract information extraction by combining the above two tasks. Then, the model is optimized and attention mechanism is adopted to further explore the correlation. Additionally, an Attention-based dynamic multitask-learning model is built. Finally, based on the above two methods, adynamic multitask learning model with lexical knowledge is proposed for the complex semantic environment in contract texts. The experimental results show that the method can fully capture the shared features among tasks and yield better information extraction results than the single-task model. It can solve the nested entity among

* 基金项目: 国家自然科学基金 (61402099, 61702093)
收稿时间: 2022-06-15; 修改时间: 2022-11-03, 2023-01-20; 采用时间: 2023-02-08; jos 在线出版时间: 2023-08-23
CNKI 网络首发时间: 2023-08-28

elements and clauses in contract texts, and realize the joint information extraction of contract elements and clauses. In addition, to verify the robustness of the proposed method, this study conducts experiments on public datasets in various fields, and the results show that the proposed method is superior to baseline methods.

Key words: multitask learning; contract text; joint information extraction; attention mechanism; nested entity

近年来, 随着我国经济实力大幅增长, 各类商务往来日益增加, 每天都会签订数以万计的合同协议, 从而产生大量与商业流程相关的合同文件, 如买卖、建设、租赁等. 这些文件包含重要的信息, 例如合同名称及编号、生效条件、合同条款等. 更重要的是, 在贸易谈判或协议修改时, 有关各方需要仔细核对合同中记录的所有条款. 因此, 从合同中提取信息已成为日常业务的重要组成部分. 但传统的合同人工审查机制需要耗费大量的人力、物力和财力, 而随着自然语言处理 (natural language processing, NLP) 的发展, 利用该技术能够从合同中高效、准确地提取所需要的信息, 提高办公效率.

合同文本信息抽取作为合同智能审查的重要组成部分, 主要分为要素抽取和条款抽取两部分: 前者负责识别合同中的要素信息, 包括合同基本信息、当事人双方信息、支付信息等; 后者负责对合同中的条款进行定位、分类并提取, 例如违约条款、解除条款等. 由于合同通常以篇章的形式出现且其中混杂着大量的要素和条款, 二者面临以下问题和挑战.

(1) 要素与条款、要素与要素之间可能存在嵌套问题, 例如支付条款内可能包含税率、总价款信息等, 总份数、甲方份数和乙方份数可能重合等.

(2) 要素和条款具有语义关联度高和字符长度差异大的两大特点. 例如质量保证条款、验收条款、解除条款中都可能存在与要素标的物的质量相关的语义信息. 当前的信息抽取研究无法很好地解决该问题[1].

(3) 相比于通用领域的信息内容, 合同要素的粒度更细、更复杂, 导致分词难度更大. 例如, 通用领域信息抽取会将“买受人 X 公司和出卖人 X 机构”中的“X 公司”和“X 机构”同时作为当事人名称提取, 而合同要素抽取会将“X 公司”和“X 机构”分别识别为“当事人名称 (甲方)”和“当事人名称 (乙方)”. 因此现有的分词方法难以达到应用要求. 当词向量作为输入时, 模型难以缓解误差传播问题; 当字符向量作为输入时, 模型难以解决一词多义的问题[2].

针对上述问题, 本文将要素抽取看作命名实体识别任务、条款抽取看作短文本分类任务, 并基于二者的高相关性, 引入多任务学习方法对合同信息进行联合抽取. 该方法的基本思想在于能够通过单一模型同时完成多项任务, 并从多个任务间的共享信息中提高任务学习效率[3,4]. 因此, 本文以买卖合同为例, 构建了合同要素与条款信息数据集; 同时将要素抽取作为主任务、条款抽取作为辅助任务, 提出了 3 种基于多任务学习的合同信息联合抽取模型, 分别是基本多任务学习模型、基于 Attention 机制的动态多任务学习模型以及基于前两者构建的引入词汇知识的动态多任务学习模型. 实验结果表明, 相比于单任务模型, 本文构建的多任务学习模型取得了结果的提升. 综上, 本文主要有以下 3 点贡献.

(1) 针对合同信息特点, 本文提出了一种新的多任务学习模型, 该模型能够捕捉篇章级合同中的上下文特征, 有效弥补了实体嵌套带来的信息提取错误和关键信息丢失.

(2) 针对合同信息抽取任务的特征, 本文在其中运用多任务学习思想, 联合抽取要素和条款, 很好地减弱了多模型多任务对抽取结果的不利影响.

(3) 针对篇章级合同的特性, 本文在多任务学习模型中引入词汇知识, 有效学习句法特征和语义信息, 解决合同文本内容易混淆的问题, 提升了联合抽取的效果.

1 相关工作

1.1 命名实体识别

在通用领域中, 命名实体识别 (named entity recognition, NER) 的目的是识别文本中的人员、位置、机构等名称[5]. 在特定领域中, NER 能抽取出特殊类型的实体, 如医学领域常用的疾病和药物[6]. NER 不仅作为信息抽取的重要组成部分, 还在诸多 NLP 下游任务中扮演关键角色, 如信息检索[7]、知识图谱构建[8]、问答系统[9]等. 当前研

究者主要着力于如何更好地利用词信息[2,10,11]. Ma 等人[10]借鉴 Lattice-LSTM[2], 在嵌入层引入词汇集合, 减少信息损失, 在诸多中文数据集上有效提升结果; Wu 等人[11]将汉字特有的偏旁部首作为特征引入 Transformer[12]结构中, 目前取得了较好的效果.

目前合同领域的 NER 研究主要集中在外文范畴, Chalkidis 等人[13]利用 75 万份合同训练出合同领域的 Word2Vec[14]词向量, 其结果优于通用领域的 GloVe[15]词向量, 表明特定领域的 NER 需要遵循本领域的特征; Chalkidis 等人[16]在其基础上, 加入 Transformer[17]结构作为对比, 结果表明合同文本中实体对其上下文高度敏感, 而缺乏时序特征的 Transformer 结构会极大影响模型识别的效果.

综上, 当前工作缺乏针对中文合同的 NER 研究, 对于合同文本特点的挖掘尚不充分, 造成合同领域的实体定义难以遵循本领域特征, 该定义方式容易误导模型, 混淆实体的概念, 如无法准确区分当事人双方信息. 此外, 现有合同领域的 NER 也没有引入词级特征信息的相关研究.

1.2 合同信息抽取

合同信息抽取的目标是从合同文件中识别具有法律效力的基本合同元素, 如签订双方的信息、金额、权力条款[18,19]. 早期的合同信息抽取方法主要是基于规则或传统机器学习的方法. Chalkidis 等人[20]通过基于人工特征工程的支持向量机提取 11 种类型的合同元素.

近期的合同信息抽取方法主要通过引入深度神经网络以提升识别效果. Chalkidis 等人[21]探索了合同要素任务的深度学习相关方法, 并构建一个不需要人工编写规则的 BiLSTM 结构. Sun 等人[22]定义 7 个特定语义的条款类别, 并构建了一种任务特定的池化层来识别合同中的嵌套要素. Wang 等人[23]探索了要素和条款之间的关系, 设计了一种双向反馈条款要素关系网络 (bi-directional feedback clause-element relation network, Bi-FLEET), 有效提升合同要素抽取的精确度.

尽管合同信息抽取技术取得了长足的进步, 但由于合同文档中复杂的上下文环境, 当前合同信息抽取对于合同内部语义特征的学习尚待挖掘.

1.3 多任务学习

多任务学习 (multitask learning, MTL) 通过训练相关联任务捕捉蕴含的领域内特定信息, 改进模型的泛化能力[24]. 随着深度学习时代的到来, MTL 转化为能够从多任务监督信息中学习共享表示的网络. 在单任务中, 每个任务通常由其自身的网络独立解决, 而与之情况相比, MTL 网络拥有以下优势: (1) 网络中的共享层能够极大地减少内存的使用率, 提高空间使用率; (2) 该结构设计避免重复计算共享编码层中的特征向量, 加快每一任务的推理速度; (3) 存在关联性的任务之间既能共享信息, 也能互补信息, 甚至能互为对方的正则化项, 避免一方任务出现过拟合现象, 提高模型的泛化能力[25].

当前 MTL 的架构被广泛应用于计算机视觉领域[26,27]和自然语言处理领域[28−31]. Singh 等人[28]在客户服务中发现用户投诉和满意度的联系, 将二者转化为投诉识别和情感分析任务, 运用多任务学习方法提升投诉识别的精确率. Tong 等人[31]将多任务学习应用于医疗命名实体识别任务, 引入句子级二分类、句子级多分类和字符级多标签分类 3 种辅助任务作为多粒度信息, 有效缓解了医疗语料缺乏的难题, 在低资源条件下表现更好. Wang 等人[30]将多分类序列标注作为辅助任务, 引入粗粒度的实体信息, 设计了一个跨文档的 NER 模型, 在多个数据集上的结果优于句子级和文档级的 NER 模型. 此外, 针对特定领域, 王卓越等人[32]、李青青等人[33]、葛海柱等人[34]将 MTL 分别应用于司法领域、医学领域、语言学领域, 实验结果表明 MTL 可增强特定领域任务的效果.

整体而言, 应用多任务学习的信息抽取已引起研究者们的关注, 通过引入多分类辅助任务, 增加粒度多样的信息, 提升实验结果. 但当前研究对辅助任务本身的效果关注较少. 此外, 篇章中的短文本信息可作为实体的扩展描述, 目前模型缺少对此类信息的应用.

2 多任务合同信息联合抽取模型

多项研究表明构建多分类辅助任务提升信息抽取主任务效果的多任务学习方法是有效的[30−32]. 本文构建了一

种基于多任务学习的模型以联合抽取合同的要素和条款信息, 同时对条款进行分类. 此外, 本文向已有的输入空间嵌入词汇知识, 进一步提升多任务学习模型的效果. 本文提出的基本框架如图 1 所示, 将在本节中详细介绍其结构和改进部分.

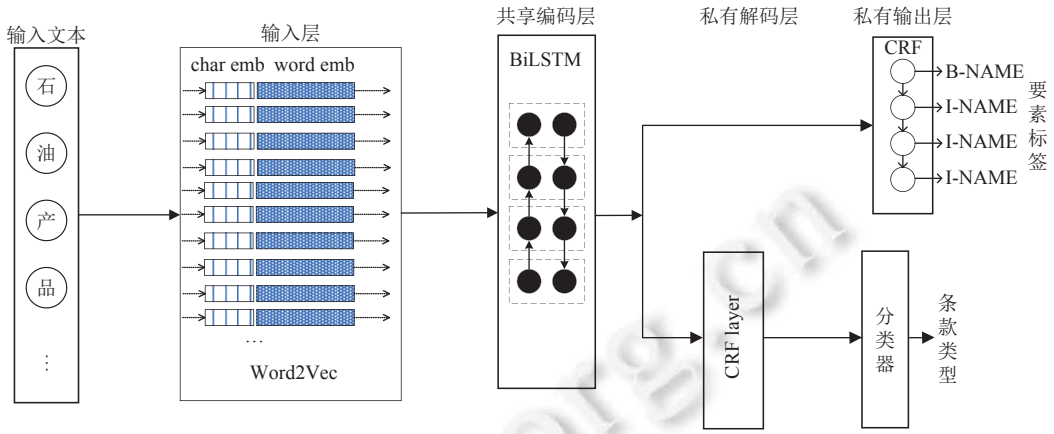

图 1 基本多任务学习模型

2.1 输入层

对于输入的语句 S, 预先训练过的词向量有助于捕获合同文本的有序性以及词与词之间的共现信息, 进而获取句法结构和语义信息. 本文沿用了 Lattice-LSTM[2]所采用的词向量 (https://github.com/jiesutd/LatticeLSTM), 它是通过 Word2Vec[14]方法训练 Giga-Word 中文语料得到的. 模型通过输入层将文本序列转化为嵌入编码序列, 作为共享编码层的输入, 便于从句子中获取上下文信息, 整个过程可简化为公式 (1).

$$X = \text{Word2Vec}(S;\theta_{\text{Word2Vec}}) \tag{1}$$

其中, X 为输入层输出的上下文语义表示, 且 $X \in \mathbb{R}^{N\times d}$, d 为嵌入向量维度, θ_{Word2Vec} 为 Word2Vec 方法的相关参数.

2.2 共享编码层

共享编码层主要负责为要素抽取任务和条款抽取任务生成共享上下文表示, 由 BiLSTM 层构成. 输入层的单词向量通过双向的 LSTM 结构, 该结构由前向和后向两个方向递归保存上下文语义信息, 并生成语句 S 中每个字符 x_t^c 的隐藏表示 h_t^c, 具体流程可简化为:

$$\overrightarrow{h_t^c} = \overrightarrow{\text{BiLSTM}}(\overrightarrow{h_{t-1}^c}, x_t^c, \theta_{\text{BiLSTM}}) \tag{2}$$

$$\overleftarrow{h_t^c} = \overleftarrow{\text{BiLSTM}}(\overleftarrow{h_{t+1}^c}, x_t^c, \theta_{\text{BiLSTM}}) \tag{3}$$

$$h_t^c = \overrightarrow{h_t^c} \oplus \overleftarrow{h_t^c} \tag{4}$$

其中, $\overrightarrow{h_t^c}$ 和 $\overleftarrow{h_t^c}$ 分别为前向和后向 BiLSTM 在当前时刻 t 的隐藏表示; θ_{BiLSTM} 为 BiLSTM 层的可训练参数; $\oplus$ 为向量拼接操作.

2.3 注意力模块

在多任务学习过程中, 不同任务对于共享编码层特征表示的应用存在差异, 且不同的辅助任务对于主任务的影响也不尽相同, 因此本文考虑采用一种动态融合的方式, 从共享特征表示中筛选出适配于不同任务的部分, 并为不同任务的特征表示分配有益于整体抽取效果的权重. 注意力机制 (Attention) 可以胜任, 其能够动态地捕捉输入编码的关联特征, 并对输入序列进行权重分配, 为关联性更高的词分配更高的值[35]. 如图 2 所示, 在本文提出的模型中, 主任务和辅助任务共享同一个 BiLSTM 编码层, 由 h_t^c 表示, 被输入到 Attention 模块中, 计算如公式 (5)–公式 (7) 所示.

$$m_t^c = \tanh(W_w h_t^c + a_w) \tag{5}$$

$$a_t^c = \frac{\exp((m_t^c)^\top m_w)}{\sum_t \exp((m_t^c)^\top m_w)} \tag{6}$$

$$s_i = \sum_t (a_t^c h_t^c) \tag{7}$$

其中, W_w 为可训练参数; m_t^c 为 h_t^c 的隐藏表示; m_w 为上下文词向量, 表示当前单词的“语境”, Attention 模块随机初始化 m_w, 并在训练过程中不断获取该向量[36]; a_t^c 为当前单词的注意力权重; s_i 为注意力模块输出的句子向量, 表示基于不同权重的单词加权和. 经过 Attention 模块, 模型将共享 BiLSTM 表示传递到两任务特定的私有层进一步学习.

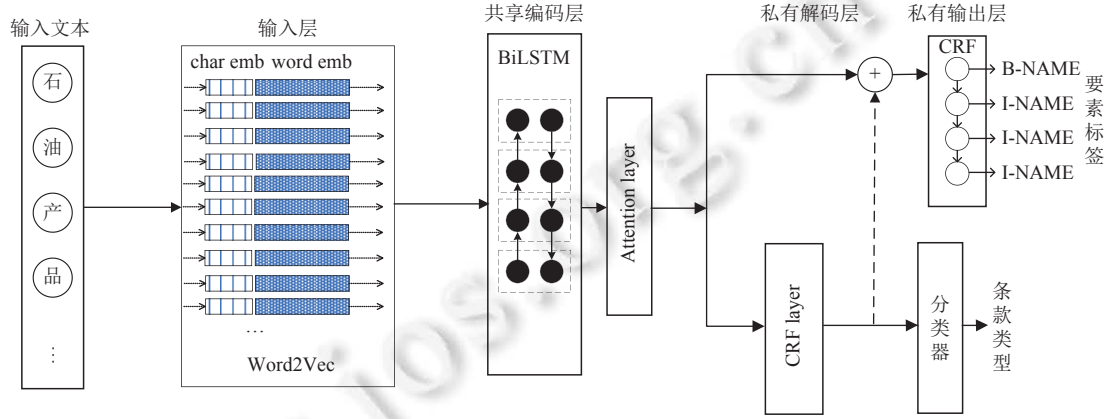

图 2 基于 Attention 机制的动态多任务学习模型

2.4 辅助任务

在单任务的要素抽取中, 尽管 NER 模型取得了较为理想的精度, 但仍存在一些预测错误, 如预测非实体为实体. 在过往研究中, NER 模型可以通过添加多种特征来提升识别精确程度[2,10,11]. 而多任务学习通过设计另一项任务来提高抽取效果已被证明是可行的[30−32,34]. 因此, 本文根据合同文本中条款较多且条款与要素间联系紧密的特点, 设计一项辅助任务, 即条款抽取任务. 为了实现该任务, 本文先通过序列标注任务将粗粒度的条款从合同文本中提取出来, 然后通过短文本分类将不同条款区分开来, 最后完成各类条款的抽取. 其中, 对于条款的序列标注, 本文参考文献 [30] 的工作, 将具体的条款类型从“BIO”序列标注方案中剔除, 仅保留“B”“I”和“O”标签, 例如, 原先表示“违约条款”的“B-BREACH”和“I-BREACH”被“B”和“I”取代. 因此, 该序列标注任务只有 3 种标签, 并且在标记策略上与要素抽取高度相关, 双方互为正则化项, 在一定程度上防止整体模型抽取的过拟合现象. 如图 1 所示, 经过共享编码层和序列标注层的编码表示, 传递到条款名称分类器中进行条款的短文本分类, 最终完成条款抽取辅助任务.

2.5 多任务学习模型

多任务学习利用相关任务间的共享特征不仅促进特征学习, 而且缩短测试过程的计算时间[37]. 具体而言, 包含一组任务的多任务学习框架比指定任务的单任务框架能提供更为精确的输出结果. 因此, 本文提出了一个多任务学习框架来解决合同联合抽取的双任务问题.

2.5.1 基本多任务学习模型

如图 1 所示, 在基本多任务学习模型 (primary multitask-learning model, PMLM) 中, 模型包括全部共享的输入层和编码层以及两个任务特定的解码层. 对于要素抽取任务, 解码层采用条件随机场 (conditional random field, CRF)[38]结构获取要素的标签; 对于条款抽取任务, 解码层先采用 CRF 结构定位条款位置, 再通过池化层和分类层对已提取的条款进行条款名称的分类. PMLM 假设全部任务无差别地共享同一特征表示, 然而共享的特征表示对于不同任务的影响是不同的, 所以 PMLM 难以处理任务间的差异性.

2.5.2 基于 Attention 机制的动态多任务学习模型

如图 2 所示, 基于 Attention 机制的动态多任务学习模型 (attention-based dynamic multitask-learning model, A-DMLM) 与基本多任务学习模型类似, 均包括共享编码层和私有解码层, 所不同的是, A-DMLM 在共享编码层后添加了 Attention 模块, 该模块能够辅助模型处理两个任务间的差异性, 将共享特征表示动态地分配给不同任务.

此外, 在基本多任务学习模型, 尽管模型为两个任务赋予了适配于当前任务的解码层, 但双方的结果并没有互相约束. 而且对于篇章级的合同文本, 条款分布在要素的上下文中, 二者紧密关联, 显然该模型并没有很好地结合此类特征. A-DMLM 针对该问题, 将条款抽取任务中序列标注模块的输出, 即粗粒度的条款信息, 作为要素抽取任务中解码层的输入, 得到最终的要素标签预测. 因此该模型能够有效地利用条款短文本信息扩展要素实体的描述.

2.5.3 基于词汇知识的动态多任务学习模型

尽管 A-DMLM 具备了联合抽取的能力, 但该模型难以有效提取合同这一类长文本内的语义信息. 为了进一步提高多任务联合抽取的效果, 本文将多种词汇知识引入 A-DMLM 中, 构建出引入词汇知识的动态多任务学习模型 (lexicon-attention-based dynamic multitask-learning model, LA-DMLM). 如图 3 所示, LA-DMLM 与 A-DMLM 主要不同在于模型在输入层添加了词汇知识模块, 该模块包含了多字词向量信息[2]和词汇位置信息[10].

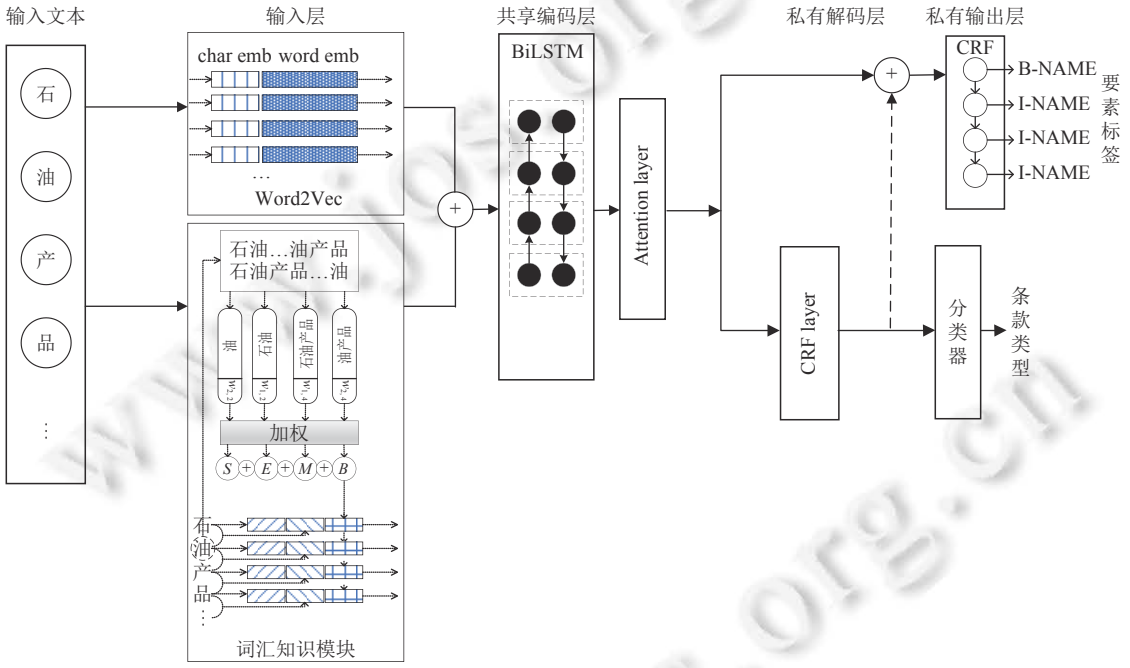

图 3 基于词汇知识的动态多任务学习模型

本文首先采用“*BMES*”的 4 位序列表示法来表示词汇位置信息. 字符 c_i 将所有与之匹配的词汇分成 4 类, 即“*B*”“*M*”“*E*”和“*S*”集合. 其中, 集合“*B*”表示当前字符位于所匹配词汇的起始部分; 集合“*M*”表示当前字符位于所匹配词汇的中间部分; 集合“*E*”表示当前字符位于所匹配词的结尾部分; 集合“*S*”表示当前字符为一个单独的字词. 该集合由“*B*”“*M*”“*E*”“*S*”标签进行标识, 如图 4 所示. 对于输入序列 $s=\{c_1,c_2,\ldots,c_m\}$ 中的字符 c_i, 4 类集合的构造方法如公式 (8)–公式 (11) 所示.

$$B(c_i)=\{w_{i,k},\forall w_{i,k}\in L,i<k\leqslant m\} \tag{8}$$

$$M(c_i)=\{w_{j,k},\forall w_{i,k}\in L,1\leqslant j<i<k\leqslant m\} \tag{9}$$

$$E(c_i)=\{w_{j,i},\forall w_{j,i}\in L,1\leqslant j<i\} \tag{10}$$

$$S(c_i)=\{c_i,\exists c_i\in L\} \tag{11}$$

其中, L 表示本文所使用的词汇表. 此外, 若某一类集合为空, 则将其置为特殊值“Null”.

其次, 本文通过词频加权的方法将集合归一化为固定维度的向量, 提高共享编码层的运行效率. 假设 $z(w)$ 为词汇 w 在数据集中出现的频率, 集合的加权表示如公式 (12), 公式 (13) 所示.

$$v^s(K)=\frac{4}{Z}\sum_{w\in S}z(w)e^w(w) \tag{12}$$

且

$$Z=\sum_{w\in B\cup M\cup E\cup S}z(w) \tag{13}$$

其中, K 表示词汇集合; e^w 表示词嵌入映射.

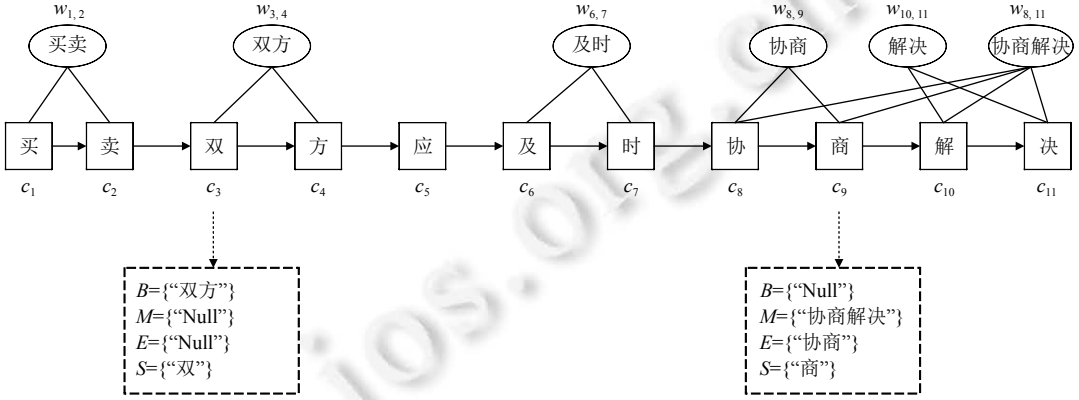

图 4 “BMES”结构图

最后, 模型将词汇位置与多字词向量两种信息进行拼接, 具体过程如下.

$$e^s(B,M,E,S)=[v^s(B);v^s(M);v^s(E);v^s(S)] \tag{14}$$

$$x_i^l=[e^u(c_i);e^b(c_i,c_{i+1})] \tag{15}$$

$$x^c\leftarrow[x^l;e^s(B,M,E,S)] \tag{16}$$

其中, $v^s(\cdot)$ 通过公式 (12) 计算得到; x_i^l 为当前字符 c_i 的多字词向量; e^u 表示一元字符向量嵌入; e^b 表示二元词向量嵌入. LA-DMLM 利用多样的词汇知识有助于处理篇章级合同文本联合抽取的挑战, 最大限度地保留关键语义信息, 进一步提升联合抽取的效率.

2.6 损失函数

本文提出的模型基于多任务学习的思想, 因此, 在联合抽取的过程中, 为了平衡两个任务, 最终的损失函数由要素抽取任务的损失 $Loss_{\text{Element}}$ 和条款抽取的损失 $Loss_{\text{Clause}}$ 加和组成, 如公式 (17) 所示.

$$Loss_{\text{Total}}=Loss_{\text{Element}}+Loss_{\text{Clause}} \tag{17}$$

具体而言, 要素抽取任务采用负对数似然函数来计算其损失, 定义如公式 (18) 所示.

$$Loss_{\text{Element}}=-\log(P(\hat{y}|x)) \tag{18}$$

其中, x 为输入的训练样本序列; $\hat{y}$ 为输入序列 x 对应的真实标签序列, 二者的序列长度一致; $P(\hat{y}|x)$ 的计算过程如公式 (19)–公式 (20) 所示. 假设输入序列 x 的长度为 l, 对于每个输入序列 $x=(x_1,x_2,x_3,\ldots,x_l)$, 都有与之对应的预测标签序列 $y=(y_1,y_2,y_3,\ldots,y_l)$, 则其得分 $score(x,y)$ 如式 (19) 所示.

$$score(x,y)=\sum_{i=0}^{l}A_{y_i,y_{i+1}}+\sum_{i=1}^{l}P_{i,y_i} \tag{19}$$

其中, $A_{y_i,y_{i+1}}$ 表示标签 y_i 到标签 y_{i+1} 的转移得分; P_{i,y_i} 表示当前字符 x_i 被标注为标签 y_i 的得分. 标记序列 y 的条件概率 $P(y|x)$ 的计算过程如公式 (20) 所示.

$$P(y|x)=\frac{\exp(score(x,y))}{\sum_{y'\in Y_x}\exp(score(x,y'))} \tag{20}$$

其中, Y_x 表示输入序列 x 对应的所有可能的标签序列. 在训练过程中, 为了使真实标签序列 $\hat{y}$ 的条件概率最大化, 模型对条件概率 $P(\hat{y}|x)$ 取负对数, 将其作为损失函数 $Loss_{\text{Element}}$, 如公式 (18) 所示.

此外, 条款抽取任务作为多分类辅助任务, 采用交叉熵损失函数进行评估, 如公式 (21) 所示.

$$Loss_{\text{Clause}}=-\sum_{i=1}^{l}\hat{y}_i\log(y_i) \tag{21}$$

其中, $\hat{y}_i$ 为真实分布; y_i 为当前字符 x_i 的预测概率分布.

3 实验结果及分析

3.1 数据集及实验设置

本文从真实业务场景和网络公开数据中收集了 1 903 份买卖合同, 将其无内容损失地转换为可标注的 TXT 文件, 并采用开源的精灵标注助手 (http://www.jinglingbiaozhu.com) 对要素和条款进行人工标注. 为了开展有效标注, 本文参考国内外数据标注工作的研究进展[39], 组织 6 名研究生作为数据标注人员, 并邀请 3 名业内专家, 对标注规范的修订以及标注数据的抽样检查提供建议和支持. 数据标注规范根据《中华人民共和国合同法》、实际业务需求以及专家团队的专业知识来制订. 标注过程分为 3 个阶段: 第 1 阶段由标注人员对部分合同进行初标注, 再经基准模型校验和专家团队的抽样检查, 对标注规范进行修订, 及时对标注人员进行培训; 第 2 阶段根据修订的规范对已标注的合同进行检验和校正, 重复第 1 阶段的校验流程, 动态调整标注规范, 形成良性循环; 第 3 阶段根据前两个阶段积累形成的反馈机制, 对所有合同数据进行标注和抽样检查, 按时完成合同标注工作. 此外, 为保证标注的准确性, 本文于数据标注过程中对标注内容进行定期抽样检查, 正确率稳定在 95% 以上. 合同数据标注工作共持续约 9 个月, 最终构建出合同实体数据集 ContractCorpus. 其中, 条款类型包括包装条款、验收条款、质量保证条款、支付条款、违约条款、争议解决条款、不可抗力条款、变更条款、解除条款、终止条款、保密条款和生效条件. 要素信息如表 1 所示, 要素类别分布不平衡, 其中, “发票类型”“银行账户 (甲方)”“传真 (乙方)”等要素类型的样本数量较少.

表 1 要素信息说明

Type	Tag	Number	Type	Tag	Number
合同编号	NUM	658	住所地 (甲方)	JWHERE	584
合同名称	NAME	1 099	住所地 (乙方)	YWHERE	570
签订日期	QTIME	749	法定代表人 (甲方)	JFN	256
签订地点	QPOS	1 168	法定代表人 (乙方)	YFN	673
合同开始时间	SHT	188	联系电话 (甲方)	JTELE	1 028
合同结束时间	EHT	453	联系电话 (乙方)	YTELE	1 041
总份数	AFEN	1 215	传真(甲方)	JCZ	209
甲方份数	JFEN	1 185	传真 (乙方)	YCZ	188
乙方份数	YFEN	1 179	邮政编码 (甲方)	JPOC	199
当事人名称 (甲方)	JNAME	1 321	邮政编码 (乙方)	YPOC	111
当事人名称 (乙方)	YNAME	1 261	标的物名称	STH	1 050
开户银行 (甲方)	JBANK	472	总价款 (大写)	LMON	1 065
开户银行 (乙方)	YBANK	1 331	总价款 (小写)	BMON	1 035
银行账户 (甲方)	JBA	92	是否含税	YNTAX	870
银行账户 (乙方)	YBA	156	税率	TAX	837
银行账号 (甲方)	JBN	379	发票类型	FPLX	27
银行账号 (乙方)	YBN	1 258	争议解决方式	ZYJJFS	903

数据集按 8:1:1 的比例随机分成训练集、验证集和测试集, 每行由两列组成, 两列间用制表符分开, 第 1 列为文本序列, 第 2 列为 BIO 格式的标签序列, 句段边界用空行标记. 需要注意的是, 条款信息的标签序列格式为“B”和“I”, 而要素信息的标签序列格式为“B-Tag”和“I-Tag”.

若将要素和条款分别看作细粒度和粗粒度的实体信息, 经本文统计, 数据集包含 214 105 个句子、10 500 069 个字符, 118 930 个实体, 句子长度 100 字以内占比 92.97%, 100 至 200 字占比 6.36%, 200 字至 300 字占比 0.42%, 300 字以上占比 0.24%, 无关标签占比约 69.1%, 含有实体的句子占比约 55.6%.

此外, 本文选取了公开数据集 Weibo[40](https://github.com/OYE93/Chinese-NLP-Corpus/tree/master/NER/Weibo) 和 Resume[2](https://github.com/jiesutd/LatticeLSTM), 数据集统计如表 2 所示, 其中 Weibo 从新浪微博中筛选出 1890 条微博信息, 进行标注形成社交领域实体数据集; Resume 从网站中爬取 1 027 份简历摘要, 通过人工标注构建实体数据集. 本文还选取了医疗领域中开源的中文电子病历数据集 CCKS17_CNER (https://www.biendata.xyz/competition/CCKS2017_2/data/), 该数据集从一组电子病历文档中标注出与医疗相关的实体, 语料统计如表 3 所示. 以上数据集的具体的实验结果及分析见第 3.4 节.

表 2 实验数据集统计

数据集	类型	Train	Dev	Test
ContractCorpus	句子	169.3k	24.8k	19.9k
	字符	8 306k	1207.3k	986.7k
Weibo	句子	1.4k	0.27k	0.27k
	字符	73.8k	14.5k	14.8k
Resume	句子	3.8k	0.46k	0.48k
	字符	124.1k	13.9k	15.1k

表 3 CCKS17_CNER 语料统计

实体类型	Train	Dev	Test
疾病	496	197	516
症状	3 784	1 497	2 257
检查	4 754	2 229	3 012
治疗	673	346	451
身体部位	4 974	2 008	2 862

在模型输入方面, 本文所使用的词向量和词汇表均来自 Lattice-LSTM[2], 其中, 词汇表包含 5.7k 个单字、291.5k 个双字词汇、278.1k 个三字词汇以及 129.1k 个其他词汇.

在模型评估方面, 本文所有实验均采用基于微平均思想的精确率 (Precision, P)、召回率 (Recall, R) 和 $F1$ 值作为模型的评价指标, 计算方式如公式 (22)–公式 (24) 所示. 当模型识别出的实体与人工标注边界和实体类型完全一致时, 算作识别完全正确. 本文模型经过多轮实验校正后, 超参数设置如表 4 所示. 本文所有模型所采用的实验环境如表 5 所示.

$$P=\frac{correct\_num}{predict\_num} \tag{22}$$

$$R=\frac{correct\_num}{true\_num} \tag{23}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{24}$$

其中, *correct_num* 表示模型正确预测的实体数量, *predict_num* 表示模型预测产生的实体总量, *true_num* 表示实际正确的实体总量, P 表示精确率, R 表示召回率.

表 4 超参数设置

Hyperparameter	Value	Hyperparameter	Value
max_sentence_length	350	hidden_dim	300
learning_rate	0.001 5	word_emb_dim	50
lr_decay	0.05	biword_emb_dim	50
clip	5	char_emb_dim	30
dropout	0.5	gaz_emb_dim	50
batch_size	16	optimizer	Adam

表 5 实验环境

参数	参数值
操作系统	Windows 10 专业版
CPU	Intel(R) Core(TM) i9-9900K CPU @ 3.6 GHz
GPU	NVIDIA GeForce RTX 3090
内存	64 GB
Python	3.6.0
PyTorch	1.8.0

3.2 有效性分析

本文提出了一种基于多任务学习的合同信息联合抽取方法, 为了验证该方法的有效性, 将其分别与单任务模型和多任务模型进行实验对比. 对于单任务模型的实验设置, 单任务的合同要素抽取模型与 LA-DMLM 对应的要素抽取部分的模型和参数设置一致, 均利用 BiLSTM 模型对输入序列进行编码、Attention 模块进行权重归一化、CRF 层进行解码, 最终产生最优标签序列, 实验结果如表 6 所示.

同理, 本文对合同条款抽取进行了相同的对比实验, 实验结果如表 7 所示.

表 6 两种不同要素抽取方法对比 (%)

模型	*P*	*R*	*F*1
single_element	81.89	83.93	82.90
multi_element	**83.42**	**84.10**	**83.76**

表 7 两种不同条款抽取方法对比 (%)

模型	*P*	*R*	*F*1
single_clause	82.70	81.16	81.92
multi_clause	**83.72**	**81.81**	**82.75**

由表 6 和表 7 可以得出, 在同等实验条件下, 多任务学习模型的结果要优于单任务学习模型的结果, 在要素抽取和条款抽取的 *F*1 值分别提高了 0.86% 和 0.83%. 两项抽取任务的实验结果说明了多任务学习的有效性, 分析原因有二: (1) 从抽取任务的角度看, 要素抽取和条款抽取均以序列标注的方法切入, 生成句子级别的标签序列, 显然二者具有关联性, 且在并行化训练的过程, 实现参数共享, 利用 Attention 模块动态更新彼此的权重, 可能令“编码-解码”网络结构更适配于多任务学习. (2) 从合同文本的角度看, 条款和要素易出现实体嵌套问题, 且二者的语义存在关联性, 单任务模型无法处理此类问题, 而多任务学习能充分利用关联特征, 完成二者的联合抽取, 尽可能避免嵌套带来的识别错误.

对于多任务模型的实验设置, 除了本文提出的 PMLM、A-DMLM 和 LA-DMLM 这 3 种模型, 本文还复现了传统的基于 BiLSTM+CRF 结构的合同信息抽取模型, 分别对要素和条款进行识别. 此外, 已有的基于多任务学习的信息抽取研究大多将序列化的多标签分类任务作为辅助手段, 通过关联任务间的相互约束来提升主任务识别的精度, 代表性的工作有 Tong 等人[31]. 由于文本特征、任务需求和评价指标不尽相同, 本文将 Tong 等人的工作 (https://github.com/zgzjdx/MT-BioNER) 进行复现, 完成要素和条款的抽取, 实验结果如表 8 所示.

表 8 多任务学习模型结果对比 (%)

模型	要素抽取 (主)			条款抽取 (辅)		
	P	*R*	*F*1	*P*	*R*	*F*1
BiLSTM+CRF	79.93	77.15	78.52	72.85	74.63	73.73
Multi_sep_tag	82.36	**84.52**	83.43	81.42	79.04	80.21
PMLM (Our)	80.17	82.01	81.08	78.82	80.46	79.63
A-DMLM (Our)	82.46	83.48	82.97	82.29	80.75	81.51
LA-DMLM (Our)	**83.42**	84.10	**83.76**	**83.72**	**81.81**	**82.75**

由表 8 中实验结果对比发现, 基于多任务学习的 PMLM 模型在要素抽取和条款抽取两项任务上的实验结果, 均优于 BiLSTM+CRF 的实验结果, *F*1 值分别提高了 2.56% 和 5.90%. 但 Multi_sep_tag、A-DMLM 和 LA-DMLM 模型在两项任务的结果均高于 PMLM 模型的结果, 主要原因在于三者不仅将序列化的标注分类作为辅助任务, 还将其输出结果作为输入主任务的粗粒度信息, 说明引入多样性的特征信息能够作为主任务实体的扩展描述, 有助于模型更有效地完成信息抽取任务. 尽管 Multi_sep_tag 在要素抽取主任务上的结果好于 A-DMLM, 但其在条款抽取辅助任务上的表现低于预期, *F*1 值略好于 PMLM 却比 A-DMLM 低, 可能原因在于 A-DMLM 的 Attention 模块能够为不同任务动态分配有效特征, 而缺乏类似模块的 Multi_sep_tag 将更多的注意力投放在主任务上, 造成条款抽取辅助任务的结果偏低.

在 A-DMLM 的基础上, 本文提出的 LA-DMLM 方法在两项任务上均取得了最佳结果, 一方面说明条款抽取任务能够很好地辅助主任务学习, 提升主任务的抽取效果, 证明了本文设计的辅助任务的合理性; 另一方面说明引

入词汇知识有利于模型更准确地定位要素和条款的边界, 进一步表明了词汇特征对于合同信息抽取的有效性.

3.3 消融分析

为了进一步验证动态分配策略和词汇知识融合的有效性, 本文对当前取得最佳效果的 LA-DMLM 模型进行消融实验, 实验结果如表 9 所示. 其中, w/o att 为模型不使用 Attention 机制对共享编码层进行动态分配的结果, w/o knowledge 为模型不使用词汇知识模块对输入序列进行编码的结果. 本文进一步对词汇知识模块内部的各项特征进行消融实验, 其中 w/o lexicon 为模型不使用词汇知识模块中的多字词向量信息进行信息抽取的结果, w/o "*BMES*"为模型不使用词汇知识模块中的词汇位置信息进行识别的结果.

表 9 消融分析结果对比 (%)

模型	要素抽取 (主)			条款抽取 (辅)		
	P	*R*	*F*1	*P*	*R*	*F*1
w/o att	82.41	84.02	83.21	81.61	80.23	80.91
w/o knowledge	80.81	82.89	81.84	82.29	80.75	81.51
w/o lexicon	81.56	82.31	81.93	82.84	**82.10**	82.47
w/o "*BMES*"	83.78	83.03	83.40	82.23	81.63	81.93
Our model	**83.42**	**84.10**	**83.76**	**83.72**	81.81	**82.75**

当缺少 Attention 模块时, 模型缺乏对共享编码层特征的有效筛选, 导致其在要素抽取和条款抽取两项任务上的 *F*1 值出现了不同程度的下降. 其中, 作为辅助的条款抽取任务的 *F*1 值下降幅度较大, 为 1.84%, 说明动态筛选共享特征对条款抽取的影响较大, 进一步验证了在多任务学习的过程中, 为不同任务分配有效的共享特征能够削弱主任务对于辅助任务的影响, 形成正向的反馈, 提升模型在要素抽取和条款抽取两项任务上的表现. 当不添加词汇知识模块时, 模型在两项任务上的 *F*1 值分别下降了 1.92% 和 1.24%, 证明了模型在输入层嵌入多样化的特征后, 丰富了句法结构信息和词级特征, 对多任务学习模型在篇章级合同文本中抽取有效的实体起到了积极的作用.

通过对比表中实验结果可知, 在词汇知识模块中, 无论缺少的是多字词向量信息还是词汇位置信息, 都会导致模型在要素抽取和条款抽取两项任务上的 *F*1 值均出现下降的问题. 其中, 缺少词汇位置信息对于条款抽取任务的影响较大, 说明"*BMES*"序列位置表示法有助于粗粒度的长文本序列化标注, 减少由要素和条款的嵌套而产生的误差. 同时, 多字词向量信息对于要素抽取任务的作用较大, 说明词汇信息可以辅助模型区分不同的实体描述, 有效地解决因实体粒度较细造成的识别错误, 从而提升要素抽取的精度.

3.4 跨领域鲁棒性分析

为了进一步说明本文提出的多任务模型的有效性, 本文分别在多个公开数据集上进行了跨领域鲁棒性实验, 数据来源包含了社交、医疗等多种领域. 本文选取了 BiLSTM+CRF[41]、Lattice-LSTM[2]、Simple Lexicon[10]和 BERT+Simple Lexicon[10]作为本节实验的基准模型, 参数设置与原论文保持一致, 实验结果如表 10 所示. 对比表中实验结果可知, LA-DMLM 在 Resume、Weibo 和 CCKS17_CNER 这 3 个数据集上的 *F*1 分别达到了 96.15%、69.90% 和 90.38%, 均优于 4 个基准模型, 说明本文提出的多任务机制在其他领域上的表现较为稳定, 模型的鲁棒性良好, 进一步验证了方法的有效性. 其中, LA-DMLM 在 Weibo 数据集上的提升并不明显, 其 *F*1 值相比于 BERT+Simple Lexicon 的 *F*1 值仅提高 0.05%.

表 10 不同领域数据集实验结果对比 (%)

模型	Resume			Weibo			CCKS17_CNER		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
BiLSTM+CRF	89.52	91.72	90.61	58.36	44.69	50.62	85.40	86.82	86.10
Lattice-LSTM	94.81	94.11	94.46	63.84	52.19	57.43	86.58	89.58	88.05
Simple Lexicon	94.76	94.29	94.53	63.47	54.50	58.64	88.16	89.34	88.75
BERT+Simple Lexicon	95.66	95.95	95.80	72.77	67.15	69.85	88.52	90.79	89.64
LA-DMLM (Our)	**95.80**	**96.50**	**96.15**	**72.84**	**67.18**	**69.90**	**88.84**	**91.98**	**90.38**

因此, 本文进一步对 LA-DMLM 应用于 Weibo 数据集的结果进行分析. 如表 11 所示, 本文统计了 Weibo 数据集中各个实体标签数量. 该数据集的实体类型分别为 PER、LOC、GPE 和 ORG, 每个实体类型分别包含命名实体 (named entity, NE) 和指代实体 (nominal mention, NM). 其中, 命名实体即为传统的 NER 任务中需要识别的实体, 指代实体主要指名词性的指代词. 以该数据集为例, PER 的命名实体 (PER.NAM) 可为“曾若彤”“胡小亭”等, 而 PER 的指代实体 (PER.NOM) 可为“哥哥们”“老师”等. 与规范严谨的新闻领域语料不同, 社交领域内的语料通常出现命名实体和指代实体混合存在的情况, 且可能有口语化的词汇夹杂其中, 例如在语料“雯子小菇凉”中, “雯子”是 PER 的命名实体, 而“小菇凉”是 PER 的指代实体, 且属于口语化的表达. 显然这种特殊的结构增大了 Weibo 数据集的实体识别难度.

LA-DMLM 在 Weibo 数据集上各类实体的识别指标精确率 (P)、召回率 (R) 和 $F1$ 值的结果如表 12 所示. 结合表 11 的实体数量分布, 本文发现模型对于数量较多的实体识别效果较好且 3 项评估指标较为均衡, 如 PER.NAM、PER.NOM 和 GPE.NAM. 一方面说明模型引入的词汇知识能够降低不规范表达带来的语义混淆, 另一方面表明模型对于地缘政治实体 (GPE.NAM) 这样的专属名词具有较高且稳定的识别结果. 尽管 ORG.NAM 实体数量与 GPE.NAM 的实体数量接近, 但 ORG.NAM 的 $F1$ 值仅为 50.10%. 可能原因有二: (1) 如图 5 所示, 由于 ORG 类型和 LOC 类型实体的表述近似且易混合出现, 造成实体识别的召回率普遍偏低, 进而影响整体的识别效果; (2) 模型可能过于关注词汇信息, 造成边界定位错误, 影响最终的识别结果.

表 11 Weibo 数据集实体标签定义及数量分布

Label	Tag	Train	Dev	Test
PER	PER.NAM	574	90	111
	PER.NOM	766	208	170
LOC	LOC.NAM	56	6	19
	LOC.NOM	51	6	9
GPE	GPE.NAM	205	26	47
ORG	ORG.NAM	183	47	39
	ORG.NOM	42	5	17

表 12 LA-DMLM 在 Weibo 数据集上各类实体的识别结果 (%)

Tag	P	R	$F1$
PER.NAM	77.20	73.67	75.40
PER.NOM	69.53	72.90	71.18
LOC.NAM	64.79	37.10	47.18
LOC.NOM	50.60	33.60	40.38
GPE.NAM	83.36	89.27	86.21
ORG.NAM	52.13	48.23	50.10
ORG.NOM	63.49	53.10	57.83

图 5 Weibo 数据集示例图

3.5 错误分析

为了说明本文提出方法的有效性, 本文以两个合同文本为例, 对典型错误类型进行分析.

例 1: 10. 争议的解决: 10.1 在本合同履行过程中发生争议时, 甲乙双方应及时协商解决. 如协商不成, 向买受人住所地法院提起诉讼. 10.2 因关联交易合同发生争议, 由双方协商解决, 协商不成的, 提交双方上级协调解决.

例 2: 买受人: 辽河油田某公司, 出卖人: 盘锦某钢材公司. 根据《中华人民共和国合同法》, 双方本着平等互利、等价有偿的原则, 经过协商一致, 资源订立本合同.

分析例 1, 正确的要素是“协商解决”“诉讼”和“协调解决”, 三者的类型均属于“争议解决方式”. 正确的条款类型为“争议解决条款”, 其内容为“在本合同履行过程中发生争议时, 甲乙双方应及时协商解决. 如协商不成, 向买受人住所地法院提起诉讼.”和“因关联交易合同发生争议, 由双方协商解决, 协商不成的, 提交双方上级协调解决.”. 显然, 要素和条款存在嵌套的问题. Multi_sep_tag 能够识别出以上 3 个要素内容, 但由于三者存在于条款之中, 造成了条款内容的截断, 该方法将此条款错误地识别为“违约条款”. 尽管 A-DMLM 能够将条款正确识别为“争议解决条款”, 但其识别的要素内容缺少“诉讼”. 相比之下, LA-DMLM 能够抽取完整的要素和条款, 并对条款进行正确的分类, 在一定程度上解决了内容嵌套的问题, 说明本文提出的 Attention 模块和基于序列化标签分类的辅助任务

对于合同信息联合抽取是有效的. 此外, 对于例 2, “辽河油田某公司”的要素类型应当为“当事人名称 (甲方)”, 而“盘锦某钢材公司”的要素类型则为“当事人名称 (乙方)”. 由于中文分词造成的定位误差, A-DMLM 和 Multi_sep_tag 错误地将甲乙方信息混淆, 而 LA-DMLM 能够正确地区分二者, 说明本文提出的引入词汇知识模块的思路有利于减少分词错误对于信息抽取的误导.

综上所述, 本文提出的基于多任务学习的 LA-DMLM 方法对于合同要素抽取和条款分类是有效的.

4 总 结

本文针对单任务模型难以处理合同文本中要素与条款嵌套的问题, 提出了 3 种多任务学习模型, 分别为 PMLM、A-DMLM 以及在此基础上引入了词汇知识的 LA-DMLM. 同时, 本文将合同条款抽取与分类作为辅助任务, 有效地提升了作为主任务的要素抽取的效果, 说明主、辅任务间具有强相关性. 值得注意的是, LA-DMLM 不仅利用 Attention 机制充分平衡主、辅任务对于整体抽取效果的影响, 而且通过词汇知识尽可能地学习篇章级合同文本的上下文语义信息, 进一步提高模型抽取的效果. 实验结果表明, 本文提出的多任务模型的结果优于主、辅任务的单任务模型. 与其他方法相比, LA-DMLM 达到了最佳的结果. 此外, 本文还在社交、医疗等领域进行了跨领域的模型鲁棒性实验, 结果表明本文提出的方法在多个公开数据集上表现均优于其他基线方法, 说明本文提出的多任务机制的鲁棒性良好.

在下一步工作中, 针对合同文本复杂的语义环境, 本文也将尝试在共享编码层引入对抗训练以削弱噪声的影响, 进一步提升模型的泛化能力和抽取效果. 此外, 本文将进一步探索合同文本的特征, 考虑添加更多的与主任务相关的辅助任务, 以研究其对多任务学习的影响.

References:

[1] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Berlin: ACL, 2016. 149–155. [doi: 10.18653/v1/P16-2025]

[2] Zhang Y, Yang J. Chinese NER using Lattice-LSTM. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: ACL, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]

[3] Tang HY, Liu JN, Zhao M, Gong XD. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In: Proc. of the 14th ACM Conf. on Recommender Systems. Virtual Event: ACM, 2020. 269–278. [doi: 10.1145/3383313.3412236]

[4] Majumder N, Poria S, Peng HY, Chhaya N, Cambria E, Gelbukh A. Sentiment and sarcasm classification with multitask learning. IEEE Intelligent Systems, 2019, 34(3): 38–43. [doi: 10.1109/MIS.2019.2904691]

[5] Deng YY, Wu CX, Wei YF, Wan ZB, Huang ZH. A survey on named entity recognition based on deep learning. Journal of Chinese Information Processing, 2021, 35(9): 30–45 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2021.09.003]

[6] Luo L, Yang ZH, Song YW, Li N, Lin HF. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning. Chinese Journal of Computers, 2020, 43(10): 1943–1957 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2020.01943]

[7] Chen YB, Xu LH, Liu K, Zeng DJ, Zhao J. Event extraction via dynamic multi-pooling convolutional neural networks. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Beijing: ACL, 2015. 167–176. [doi: 10.3115/v1/P15-1017]

[8] Hong WX, Hu ZQ, Weng Y, Zhang H, Wang Z, Guo ZX. Automated knowledge graph construction for judicial case facts. Journal of Chinese Information Processing, 2020, 34(1): 34–44 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2020.01.005]

[9] Diefenbach D, Lopez V, Singh K, Maret P. Core techniques of question answering systems over knowledge bases: A survey. Knowledge and Information Systems, 2018, 55(3): 529–569. [doi: 10.1007/s10115-017-1100-y]

[10] Ma RT, Peng ML, Zhang Q, Wei ZY, Huang XJ. Simplify the usage of lexicon in Chinese NER. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5951–5960. [doi: 10.18653/v1/2020.acl-main.528]

[11] Wu S, Song XN, Feng ZH. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). ACL, 2021. 1529–1539. [doi: 10.18653/v1/2021.acl-long.121]

[12] Li XN, Yan H, Qiu XP, Huang XJ. FLAT: Chinese NER using flat-lattice Transformer. In: Proc. of the 58th Annual Meeting of the

Association for Computational Linguistics. ACL, 2020. 6836–6842. [doi: 10.18653/v1/2020.acl-main.611]

[13] Chalkidis I, Fergadiotis M, Malakasiotis P, Androutsopoulos I. Neural contract element extraction revisited. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver, 2019. 7413–7424.

[14] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the 26th Int'l Conf. on neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.

[15] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). DohaL ACL, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[16] Chalkidis I, Fergadiotis M, Malakasiotis P, Androutsopoulos I. Neural contract element extraction revisited: Letters from sesame street. arXiv:2101.04355, 2021.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[18] Curtotti M, McCreath E. Corpus based classification of text in Australian contracts. In: Proc. of the 2010 Australasian Language Technology Association Workshop. Melbourne, 2010. 18–26.

[19] Indukuri KV, Krishna PR. Mining e-contract documents to classify clauses. In: Proc. of the 3rd Annual ACM Bangalore Conf. Bangalore: ACM, 2010. 7. [doi: 10.1145/1754288.1754295]

[20] Chalkidis I, Androutsopoulos I, Michos A. Extracting contract elements. In: Proc. of the 16th Edition of the Int'l Conf. on Articial Intelligence and Law. London: ACM, 2017. 19–28. [doi: 10.1145/3086512.3086515]

[21] Chalkidis I, Androutsopoulis I. A deep learning approach to contract element extraction. In: Wyner AZ, Casini G, eds. Frontiers in Artificial Intelligence and Applications: Vol. 302, Legal Knowledge and Information Systems. IOS Press, 2017. 155–164. [doi: 10.3233/978-1-61499-838-9-155]

[22] Sun L, Zhang K, Ji FL, Yang ZH. Toi-CNN: A solution of information extraction on Chinese insurance policy. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Industry Papers). Minneapolis: ACL, 2019. 174–181. [doi: 10.18653/v1/N19-2022]

[23] Wang ZH, Song HY, Ren ZC, Ren PJ, Chen ZM, Liu XZ, Li HS, de Rijke M. Cross-domain contract element extraction with a bi-directional feedback clause-element relation network. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Virtual Event: ACM, 2021. 1003–1012. [doi: 10.1145/3404835.3462873]

[24] Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41–75. [doi: 10.1023/A:1007379606734]

[25] Vandenhende S, Georgoulis S, Van Gansbeke W, Proesmans M, Dai DX, van Gool L. Multi-task learning for dense prediction tasks: A survey. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2022, 44(7): 3614–3633. [doi: 10.1109/TPAMI.2021.3054719]

[26] Chu Z, Mi Q, Ma W, Xu SB, Zhang XP. Keypoint-level occlusion-aware human pose estimation. Journal of Computer Research and Development, 2022, 59(12): 2760–2769 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.20210723]

[27] Liu SK, Johns E, Davison AJ. End-to-end multi-task learning with attention. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1871–1880. [doi: 10.1109/CVPR.2019.00197]

[28] Singh A, Saha S, Hasanuzzaman M, Dey K. Multitask learning for complaint identification and sentiment analysis. Cognitive Computation, 2022, 14(1): 212–227. [doi: 10.1007/s12559-021-09844-7]

[29] El-Allaly ED, Sarrouti M, En-Nahnahi N, El Alaoui SO. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. Information Processing & Management, 2021, 58(3): 102473. [doi: 10.1016/J.IPM.2020.102473]

[30] Wang DS, Fan HJ, Liu JF. Learning with joint cross-document information via multi-task learning for named entity recognition. Information Sciences, 2021, 579: 454–467. [doi: 10.1016/j.ins.2021.08.015]

[31] Tong YQ, Chen YD, Shi XD. A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. In: Proc. of the 2021 Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. ACL, 2021. 4804–4813. [doi: 10.18653/v1/2021.findings-acl.424]

[32] Wang ZY, Chen YG, Xing TJ, Sun YY, Yang L, Lin HF. Joint entity and relation extraction for multi-crime legal documents with multi-task learning. Computer Engineering and Applications, 2023, 59(2): 178–184 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2108-0344]

[33] Li QQ, Yang ZH, Luo L, Lin HF, Wang J. A multi-task learning approach to biomedical entity relation extraction. Journal of Chinese Information Processing, 2019, 33(8): 84–92 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2019.08.012]

[34] Ge HZ, Kong F. Chinese elementary discourse unit and theme-rheme joint detection based on multi-task learning. Journal of Chinese Information Processing, 2020, 34(1): 71–79 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2020.01.010]

[35] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2016.

[36] Sukhbaatar S, Szlam A, Weston J, Fergus R. End-to-end memory networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems, Vol. 2. Montreal: MIT Press, 2015. 2440–2448.

[37] Zhang Y, Yang Q. A survey on multi-task learning. IEEE Trans. on Knowledge & Data Engineering, 2022, 34(12): 5586–5609. [doi: 10.1109/TKDE.2021.3070203]

[38] Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 2001. 282–289.

[39] Cai L, Wang ST, Liu JH, Zhu YY. Survey of data annotation. Ruan Jian Xue Bao/Journal of Software, 2020, 31(2): 302–320 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5977.htm [doi: 10.13328/j.cnki.jos.005977]

[40] Peng NY, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 548–554. [doi: 10.18653/v1/D15-1064]

[41] Yang J, Zhang Y. NCRF++: An open-source neural sequence labeling toolkit. In: Proc. of the 2018 ACL System Demonstrations. Melbourne: ACL, 2018. 74–79. [doi: 10.18653/v1/P18-4013]

附中文参考文献:

[5] 邓依依, 邬昌兴, 魏永丰, 万仲保, 黄兆华. 基于深度学习的命名实体识别综述. 中文信息学报, 2021, 35(9): 30–45. [doi: 10.3969/j.issn.1003-0077.2021.09.003]

[6] 罗凌, 杨志豪, 宋雅文, 李楠, 林鸿飞. 基于笔画ELMo和多任务学习的中文电子病历命名实体识别研究. 计算机学报, 2020, 43(10): 1943–1957. [doi: 10.11897/SP.J.1016.2020.01943]

[8] 洪文兴, 胡志强, 翁洋, 张恒, 王竹, 郭志新. 面向司法案件的案情知识图谱自动构建. 中文信息学报, 2020, 34(1): 34–44. [doi: 10.3969/j.issn.1003-0077.2020.01.005]

[26] 褚真, 米庆, 马伟, 徐士彪, 张晓鹏. 部位级遮挡感知的人体姿态估计. 计算机研究与发展, 2022, 59(12): 2760–2769. [doi: 10.7544/issn1000-1239.20210723]

[32] 王卓越, 陈彦光, 邢铁军, 孙媛媛, 杨亮, 林鸿飞. 基于多任务学习的多罪名案件信息联合抽取. 计算机工程与应用, 2023, 59(2): 178–184. [doi: 10.3778/j.issn.1002-8331.2108-0344]

[33] 李青青, 杨志豪, 罗凌, 林鸿飞, 王健. 基于多任务学习的生物医学实体关系抽取. 中文信息学报, 2019, 33(8): 84–92. [doi: 10.3969/j.issn.1003-0077.2019.08.012]

[34] 葛海柱, 孔芳. 基于多任务学习的汉语基本篇章单元和主述位联合识别. 中文信息学报, 2020, 34(1): 71–79. [doi: 10.3969/j.issn.1003-0077.2020.01.010]

[39] 蔡莉, 王淑婷, 刘俊晖, 朱扬勇. 数据标注研究综述. 软件学报, 2020, 31(2): 302–320. http://www.jos.org.cn/1000-9825/5977.htm [doi: 10.13328/j.cnki.jos.005977]

王浩畅(1974－), 女, 博士, 教授, CCF 高级会员, 主要研究领域为自然语言处理, 数据挖掘, 生物信息学.

赵铁军(1962－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器翻译, 人工智能.

郑冠彧(1997－), 男, 硕士, 主要研究领域为自然语言处理, 信息抽取, 命名实体识别.