反馈机制的实体及关系联合抽取方法

马建红 李振振 朱怀忠 魏宇默
(河北工业大学人工智能与数据科学学院 天津 300401)

摘 要 实体及关系抽取是信息抽取中的两个核心任务,是构建知识图谱的重要基石。对于实体识别和关系抽取,当前通常采取人工提取特征和规则,分独立两步实现的方法,这种方法易造成数据重复预处理和错误传播。实体识别和关系抽取两个模块存在相互关联性,实体识别是进行关系抽取的基础,实体关系抽取结果又可反馈校验实体信息。因此,文中提出无须添加人工特征和引入互反馈机制的混合神经网络模型(Mufeedback-Join Model)来完成实体及其关系的联合抽取,实现实体关系对实体识别的反馈校验机制。该模型共享 Bi-LSTM 特征提取层来提取文本上下文特征,依据共享层特征引入 Attention 机制捕获关键局部特征来完成解码,再用条件随机场 CRF 完成实体序列的标注任务,融合共享层特征和实体特征,并将其输入到 CNN 模型来实现实体关系的抽取,最后计算关系抽取结果的损失值,再联合实体识别损失值反馈修正特征提取层和实体识别模型参数。将此算法应用在实体数据集上进行实验,在同等硬件环境下,该方法可以缩短的模型训练时间,提升实体及关系抽取的准确率、召回率和 F1 值,联合抽取的 F1 值整体提升了 3.91%,实体识别子模块的 F1 值平均提升了 1.34% ,关系抽取的 F1 值提升了 5.79%。实验数据说明,联合抽取模型可以实现两个子模块的合并,从而缩短数据处理时间和减少错误数据的传递;相互反馈的机制可以提升整体识别效果。

关键词 反馈机制,联合抽取,深度学习,实体识别,关系抽取
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.181102117

Entity and Relationship Joint Extraction Method of Feedback Mechanism

MA Jian-hong LI Zhen-zhen ZHU Huai-zhong WEI Zi-mo
(School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract Entity and relationship extraction are two core tasks in information extraction, and are the important cornerstone of knowledge mapping. At present, entity recognition and relationship extraction usually adopt the method of extracting features and rules manually and realizing them independently in two steps. This method is easy to cause duplicate data preprocessing and error propagation. The two modules are interrelated. Entity recognition is the basis of relationship extraction. The results of entity relationship extraction can also feedback and verify entity information. Therefore, a hybrid neural network model (Mufeedback-Join Model) without adding manual features and with mutual feedback mechanism was proposed to extract entities and their relationships jointly and realize the feedback checking mechanism of entity relationship to entity recognition. The model shares Bi-LSTM feature extraction layer to extract text context features, and introduces attention mechanism to capture key parts based on shared layer features. After decoding the feature, CRF is used to complete the entity sequence labeling task. The shared layer feature and entity feature are input into CNN model to realize entity relationship extraction. Finally, the relationship extraction result loss value is calculated, and the feature extraction layer of loss value feedback correction and the parameters of entity recognition model are combined. In the same hardware environment, this method can shorten the training time of model, improve the accuracy, recall and F1 value of entity and relationship extraction, The F1 value of the joint extraction is improved by 2.91%, the entity identification sub-module F1 is increased by 1.34% on average, and the relationship extraction F1 value is increased by 5.79%. The experimental data show that the joint extraction model can merge two sub-modules to reduce data processing time and error data transmission, and the mechanism of mutual feedback can improve the overall recognition effect.

Keywords Feedback mechanism, Joint extraction, Deep learning, Entity recognition, Relation extraction

---

到稿日期:2018-11-16 返修日期:2019-03-31 本文受河北省科技厅互联网的创新软件设计及公共应用服务平台项目(15240118D)资助。
马建红(1965-),女,博士,教授,CCF 会员,主要研究方向为软件工程、智能处理、自然语言处理、知识图谱、创新理论与方法,E-mail:m_zh2002@126.com(通信作者);**李振振**(1992-),男,硕士,主要研究方向为自然语言处理、软件工程;**朱怀忠**(1978-),男,硕士,讲师,主要研究方向为计算机应用、人工智能;**魏宇默**(1994-),男,硕士,主要研究方向为软件工程、自然语言处理。

1　引言

实体和关系抽取是为了识别出一段自由文本中提到的实体信息及实体之间隐含的语义关系。例如，给定一段文本“聚甲醛即使在低温下仍有很好的抗蠕变特性”，在这段文本中存在资源实体“聚甲醛”、属性实体“抗蠕变特性”和参数实体“很好”。其中，实体对〈聚甲醛，抗蠕变特性〉蕴含着资源和属性的关系，实体对〈抗蠕变特性，很好〉蕴含着属性和量值的关系。完成这项任务的传统方法是把实体识别和关系抽取视为两个完全独立的任务，即将其分别看作资源命名实体识别(Named Entity Recognition，NER)[1-2]任务和关系抽取(Relation Extraction，RE)[3]任务，NER 任务结束后，根据 NER 的识别结果重新提取文本特征再做实体关系抽取，这种方式被称为管道模型。

管道模型各部分的独立性强，但每一部分需要重复底层特征提取和训练工作，忽视了两个子任务之间的相互关联性。联合学习[4-7]框架是将 NER 和 RE 关联起来的有效方式。联合学习模型将两个问题归结为同一个问题，构建新的联合模型。联合的方式可以解决如下问题：1)错误传播，避免实体识别模块的错误影响到关系抽取模块的性能；2)数据重复处理，避免了两个任务都需要进行数据预处理的重复性工作。现有的联合学习模型存在 NER 和 RE 相互关联性不足的问题。

根据以上分析，本文提出了一种新的具有反馈机制的联合模型。首先引入 Attention 机制提升 NER 性能，再用两个模块的互反馈机制提升实体识别和关系抽取模块的相关性，实现实体及关系的联合抽取。本文将在第 3 节详细描述联合模型的结构和反馈机制。

2　相关研究

2.1　实体及关系管道抽取方法

现阶段 NER 和 RE 模型主要存在 3 类方法：基于规则的方法、基于人工特征[8]的机器学习方法和现阶段无人工特征的基于神经网络的方法。

1)命名实体识别

最初的 NER 主要采用的是基于规则[9]的识别，通过领域专家和语言学者人工制定有效规则来识别命名实体。此方法仅适用于简单的识别场景，对于复杂的 NER 任务，需要耗费大量的时间和精力来制定规则，且领域迁移性差。后来，学者们采用了机器学习的方法，把 NER 任务视为一种序列标注任务，将每一字或者词都标记为一个标签类别。机器学习中有多种解决序列标注任务的方法，如隐马尔可夫模型(Hidden Markov Models，HMM)[10]、最大熵马尔可夫模型(Maximum Entropy Markov Models，MEMM)、条件随机场模型(Conditional Random Fields，CRF)[11]等，这些方法需要研究者们人工提取有效的语法特征。最近，神经网络的方法被成功地运用到 NER 领域。张海楠等[12]提出自动学习字特征和词特征的混合网络模型，其不再依赖人工特征。Miwa 等[5]提出编码解码模型，在提取特征后引入 Bi-LSTM 编码，再利用长短时记忆网络(Long Short-Term Memory，LSTM)解码，最后通过 softmax 确定分类标签。Dong 等[13]提出基于 Bi-LSTM 利用 CRF 来进行中文的命名实体识别，该方法利用 CRF 能较好地捕获时间序列的特性，提高识别准确率，但在长语句语料中表现不佳。Luo 等[14]在 Bi-LSTM 中引入注意力机制(Attention)以关注长文本的局部特征，所提方法在对医疗文本的实体提取中有显著的效果提升。以上 RNN 模型都是以 Bi-LSTM 为编码模型，但是它们的解码方式存在很大差异。

2)实体关系抽取

基于人工特征的方法主要使用 NLP 工具和领域知识来获取有效的人工特征。不同的分类模型往往采用不同的特征集合，特征类型主要有 3 种[9]：1)词汇特征，包括实体词义、词性标注、相邻实体信息；2)句法特征，包括词组块、语法树；3)语义特征，包括实体类型。Kambhatla 等[8]使用最大熵(ME)模型结合特征进行分类。车万翔等[15]使用实体类别、实体位置关系、前后词信息特征，然后使用支持向量机(SVM)方法实现关系的抽取，取得了较好的效果。陈宇等[3]在上述特征之外，又划分包含关系和非包含关系、实体对根节点、实体间路径、依存动词与实体路径等句法结构信息特征。马晓军等[16]用 Bootstrapping 方法训练匹配模式，不断迭代扩充置信水平较高的关系实体，并将其加入到种子集，但此方法需要较大的知识库作为支撑。

应用在关系抽取中的神经网络模型大多是卷积神经网络(Convolutional Neural Networks，CNN)和长短时记忆网络。Socher 等[17]使用 RecNN 模型学习基于分析树的结构特征，证明了结构表示的有效性。然而，递归结构的模型在句子较长时的时间复杂度较高，因此 RecNN 模型不适用于长文本。Lai 等[18]采用 RNN 来降低时间复杂度，但单纯的 RNN 模型又存在网络偏置问题。Yan 等[19]采用双向的 LSTM 和最大池化层来解决遗忘问题，展示了 LSTM 网络在关系抽取中的有效性。考虑到单纯以词序列作为输入会导致在表达语法结构上存在缺陷，Li 等[20]采用基于树结构的 LSTM 模型并进行实验，结果表明所提模型的效果优于基于词序列的 LSTM。

2.2　实体及关系联合抽取方法

管道方式中，关系抽取模型的性能依赖于实体识别的效果，每一模块都要从最底层的数据预处理做起。

针对管道模型的不足，Dan[21]和 Yang[22]提出用联合学习的模型来优化子任务的结果和寻找全局最优的解决方案。Singh 等[23]提出图连接模型来表示两个子任务之间的各种依赖关系。Li 等[24]首次提出用一个基于人工神经网络的联合模型来预测实体及实体关系，这种模型是一种具有高效搜索功能的结构感知器。Miwa 等[4]引入一个表来表达句子中的实体和关系结构，并且提出了一个基于历史的学习模型。Miwa 等[5]又提出了一个基于 LSTM 的模型来抽取实体和关系，该模型可以有效地减少人工工作量。以上这些模型优化了底层特征提取，但对 NER 和 RE 模块的相关性重视不足。本文在之前研究的基础上构建具有互反馈机制的联合抽取模型来抑制错误传播，共享底层特征。

3 反馈式联合抽取模型的构建

实体和关系反馈式联合抽取模型的框架由共享模块、NER 模块、RE 模块构成，如图 1 所示。NER 和 RE 都需要提取句子词向量特征和句子上下文语义特征，而 Bi-LSTM 模型能够捕获上下文信息，因此两模块共享 Bi-LSTM 编码特征。在共享层特征的基础上引入 NER 模块和 RE 模块（NER 模块如图 2 所示），先对 Bi-LSTM 特征引入 Attention 机制，为关键词分配较高权重，以减少无关词的干扰。序列标注带有时序性，因此解码后用 CRF 模型完成序列标注任务，提取实体信息。实体抽取结果将进一步辅助完成实体关系抽取，在 RE 模块，首先根据实体标签计算实体特征，并将计算后的特征输入到 RE 模块，完成关系抽取任务；然后依据 RE 损失值调整 RE 模块参数，完成联合抽取模型的前馈过程。

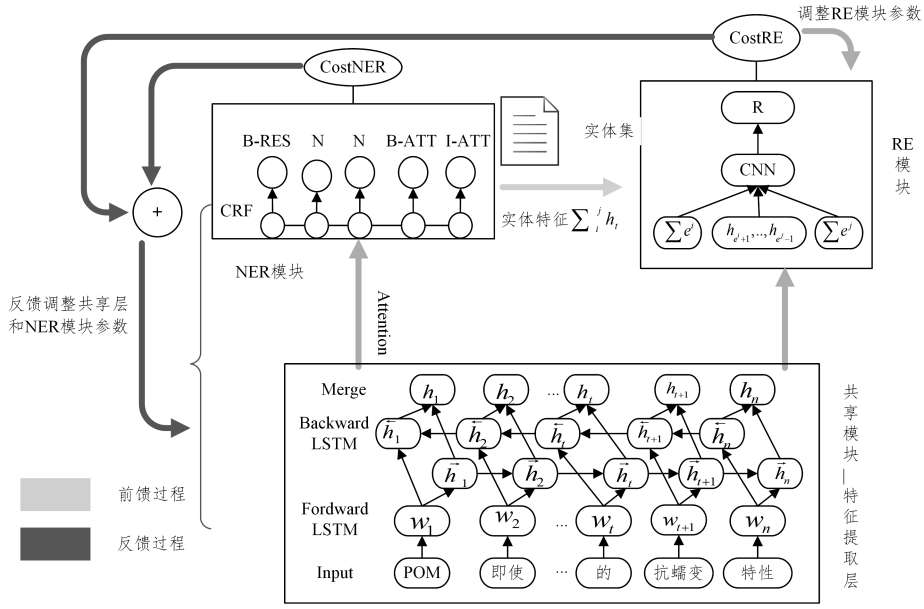


图 1 反馈式联合抽取框架图

Fig. 1 Framework of joint extraction of feedback mechanism

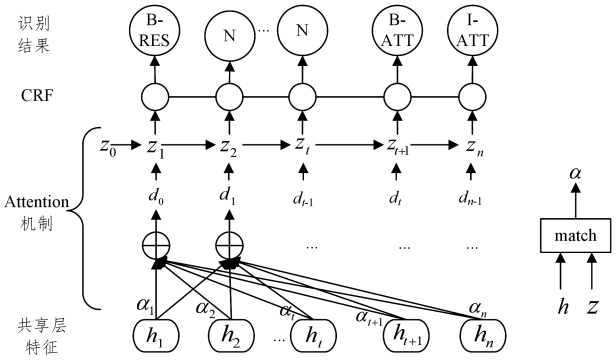


图 2 引入 Attention 机制的 NER 模型

Fig. 2 NER model with Attention mechanism

实体信息越准确，实体之间的语义关系就越容易准确确定，因此关系抽取的结果可以作为实体识别性能的校验。实体关系抽取的准确率高，说明抽取到的实体信息较为准确；反之，说明抽取到的实体错误率偏高。那么，完成实体关系抽取后将结果反馈到实体识别模型，NER 模块和共享层模块依据 RE 模块的损失值及准确率和 NER 模块的损失值重新调整特征提取层和 NER 模块参数，共享模块和实体识别模块不再单一地依赖 NER 结果，而是融入 RE 的反馈信息。反馈调节过程将由算法 1 具体描述。

3.1 联合抽取算法

联合模型抽取算法如算法 1 所示，相关符号说明如表 1 所列。算法的输入是经过数据预处理后得到的语料集 D、用于控制 NER 模块和 RE 模块对共享层参数的影响系数 α 和 β、模型最大训练迭代次数 K。算法执行完成后，将返回识别到的实体集合 PE，以及实体关系集合 PR。

算法 1 反馈式联合模型算法 MFB_JoinModel

输入：语料集 D，超参数 α 和 β，迭代次数 K

输出：预测实体集 PE，预测关系集 PR

```
1. PE=PR=∅
2. Θ=γ=0
3. for k=1 to K do
4.     for d∈D do
5.         pe←NER(d)
6.         cost_ner←cost(pe,re)
7.         for e_i∈pe do
8.             for e_j∈pe−{e_1…e_i} do
9.                 r←RC(e_i,e_j)
10.                pr←pr∪(e_i,e_j,r)
11.            end
12.        end
13.        cost_rc←Σ_{i=1}^{|rc|} cost(pr,rr)
14.        γ←γ′←adjust(cost_rc)
15.        cost←αcost_ner+βcost_rc
16.        Θ←Θ′←adjust(cost)
17.        PE←PE∪pe,PR←PR∪pr
18. end
19. return PE,PR
```

表 1　MFB_JoinModel 算法符号说明

Table 1　MFB_JoinModel algorithm symbol explanation

变量	描述
PE，PR	预测实体集合，预测关系集合
pe，pr	当前语料预测实体集，实体关系集
re，rr	当前语料真实实体集，实体关系集
Θ	共享层及 NER 模块参数矩阵
γ	RC 模块参数矩阵
$cost\_ner$	NER 模块损失值
$cost\_re$	RE 模块损失值
$cost$	NER 和 RE 联合损失值
e_i，e_j	实体特征向量
r	实体的语义关系
$cost(par1,par2)$	损失计算函数，参数 $par1$ 为预测值，$par2$ 为真实值

实体识别和关系抽取模块归结为多分类问题，因此每一部分的损失都可选用式(1)的方法来计算其多分类逻辑回归损失值。其中，$\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ 是对参数的正则化，防止出现过拟合问题，λ 为可控的惩罚因子，θ 为要训练的参数，n 是模型中要训练参数的个数，K 是分类个数，m 是训练样本数。

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}[-y_k^{(i)}\log((h_\theta(x^{(i)}))_k)-(1-y_k^{(i)})\log(1-(h_\theta(x^{(i)}))_k)]+\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \tag{1}$$

$$h_\theta(x^{(i)})=\frac{1}{1+e^{-\theta^T x^{(i)}}} \tag{2}$$

算法 1 中，第 1－2 行初始化变量；第 3 行控制模型的迭代次数；第 4 行处理数据集中的每一条数据语料；第 5 行由 NER 模型识别出输入语料中的实体词，$NER()$ 表示实体识别模型；第 6 行计算当前识别结果的损失值，$cost()$ 为损失计算函数，其计算公式如式(1)所示；第 7－12 行是关系抽取模块，根据 NER 模块识别到的实体集合对当前实体及当前实体之后的实体构成的实体对一一进行实体语义关系抽取；第 9 行由关系抽取模块抽取实体语义关系；第 10 行将识别到的语义关系添加到当前语料关系集合；第 13 行计算当前语料下所有实体语义关系识别结果的损失值；第 14 行依据损失值调整 RE 模块的参数；第 15 行计算 NER 模块和 RE 模块的加权损失和；第 16 行依据 15 行计算的损失值调整共享层和 NER 模块的参数，重新调整时不再单一依赖 NER 识别结果，引入了 RE 损失值的反馈；第 17 行将当前语料识别的实体集合和实体间的语义关系并入到模型总结果集；第 19 行返回识别到的实体集和实体语义关系集。

3.2　共享层模块

在共享模块中完成文本上下文特征的提取，图 1 中的特征提取层展示了计算词向量特征和计算词上下文特征的过程。以 one-hot 模型表示词特征时，存在维度较高和特征稀疏等问题，因此引入 Embedding 层，用基于神经网络的词向量转化工具 Word2vec 将词的 one-hot 特征转化为实数集上的低维稠密特征。Word2vec 选用 Skip-Gram 实现，在中小数据规模下，相比于 CBOW 方法，以相对较高的时间复杂度提高了词向量的准确性。因此，可以将一个句子序列表示为 $w=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_t,\cdots,\boldsymbol{w}_n\}$，其中 $\boldsymbol{w}_t\in R^d$ 表示在一句语料中第 t 个词的词向量，词向量的维数是 d 维，n 表示输入句子的长度。

在 Embedding 层之后，引入一个前向 LSTM(F-LSTM)和后向 LSTM(B-LSTM)构成一个双向的 LSTM 层(Bi-LSTM)网络。对于每一个输入的词向量，F-LSTM 将根据 $\boldsymbol{w}_1$ 到 $\boldsymbol{w}_t$ 的词上文信息编码词向量 $\boldsymbol{w}_t$，其输出结果记为 $\overrightarrow{\boldsymbol{h}_t}$。同样地，B-LSTM 将根据词 $\boldsymbol{w}_n$ 到 $\boldsymbol{w}_t$ 的下文信息编码 $\boldsymbol{w}_t$，其输出结果记为 $\overleftarrow{\boldsymbol{h}_t}$。LSTM 采取自适应的门机制，这些门控能决定记忆单元保留上一级记忆状态和提取当前输入特征的程度。LSTM 结构包含了一系列循环连接的子网络，双向 LSTM 的前向隐藏层和后向隐藏层的每一个时间步都是一个记忆块，如图 3 所示。

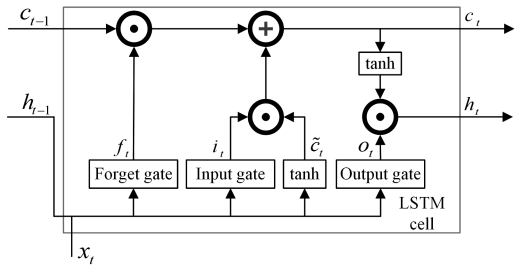


图 3　LSTM 细胞单元结构图

Fig. 3　LSTM memory block with one cell

每一个记忆块包含一个或多个自连接的记忆细胞和 3 个乘积单元，每个乘积单元分别连接输入门 i、输出门 o 和遗忘门 f。每一个 LSTM 记忆块基于前一隐藏层向量 $\boldsymbol{h}_{t-1}$、前一个细胞记忆向量 $\boldsymbol{c}_{t-1}$ 和当前输入的词向量来计算当前隐藏层向量 $\boldsymbol{h}_t$，可以记为 $\overrightarrow{\boldsymbol{h}_t}=lstm(\overrightarrow{\boldsymbol{h}_{t-1}},\overrightarrow{\boldsymbol{c}_{t-1}},\boldsymbol{w}_t)$ 和 $\overleftarrow{\boldsymbol{h}_t}=lstm(\overleftarrow{\boldsymbol{h}_{t-1}},\overleftarrow{\boldsymbol{c}_{t-1}},\boldsymbol{w}_t)$，下标 t 表示时间序列位置，当前输入为 x_t。

$$i_t=\sigma(w_i[h_{t-1},x_t]+b_i) \tag{3}$$

$$f_t=\sigma(w_f[h_{t-1},x_t]+b_f) \tag{4}$$

$$o_t=\sigma(w_o[h_{t-1},x_t]+b_o) \tag{5}$$

$$\tilde{c}_t=\tanh(w_c[h_{t-1},x_t]+b_c) \tag{6}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ\tilde{c}_t \tag{7}$$

$$h_t=o_t\circ\tanh(c_t) \tag{8}$$

其中，i，f 和 o 表示输入门、遗忘门和输出门，b 表示偏置单元，c 是记忆细胞，$\circ$ 表示每个向量的数乘，$w_{(\cdot)}$ 表示神经元连接参数。最后联合向量 $\overrightarrow{\boldsymbol{h}_t}$ 和 $\overleftarrow{\boldsymbol{h}_t}$ 表示第 t 个词的特征，记为 $\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}_t},\overleftarrow{\boldsymbol{h}_t}]$。

3.3　命名实体识别(NER)模块

每一个词将被指派到一个实体标签，每个标签都用相同的编码模式，即 BI(Begin，Inside)，每一个标签表示实体中一个词的位置信息。这里主要有 3 种实体：资源实体(RES)，属性实体(ATT)和量值实体(VAL)。每一个实体都包含开始和中间标记，非实体词用 O 标记，标签如表 2 所列。

表 2　实体标签

Table 2　Entity label

	RES	ATT	VAL
B	B-RES	B-ATT	B-VAL
I	I-RES	I-ATT	I-VAL

共享层计算每一个词的上下文特征 h_t，$H=(h_1,h_2,\cdots,$

$h_t,\cdots,h_m$)作为命名实体识别模块的输入。在对编码层的信息解码时引入 Attention 机制,Attention 机制的概念最早出现在图像处理领域,用于捕获图像的局部信息,现已成功应用在 NLP 领域。无 Attention 机制的模型中每一个词对产生实体标注结果的影响是相同的。然而,在实际语料中,有些词对识别结果有较大影响,例如对于"特性"一词,在该词附近往往存在着一个属性实体。每个词对分类结果的影响程度不同,Attention 机制是要为每个词定义不同的影响系数,在判定某一时间片的输出时,更加依赖某些关键局部特征,排除了干扰因素。

Attention 模型本质上是一种相似性度量,当前输入与目标状态越相似,当前输入的权重就会越大,当前的输出也就越依赖于当前的输入。如图 2 所示,Attention 机制下,由 match 模块计算当前输入和输出的匹配度,match 是一个计算模型,可以是 h_t 和 z_{t-1}(其中 z_0 为初始化向量)的余弦相似度,也可用一层 BP 神经网络自动学习匹配度[25]。然后将当前的输出和每一个输入做一次 match 计算,分别得到当前输出和所有输入的匹配度 α_i;再使用 softmax 进行归一化,使其输出时所有权重之和为 1,即 $\sum_{i=1}^{n}\alpha_i=1$。计算出其加权向量和以后,将其作为下一层的输入。其中:

$$d_t=\sum_i \alpha_i^t h^i \tag{9}$$

在实体标注部分,对于给定的长度为 m 的序列 H,假设标注结果 $Y=(y_1,\cdots,y_t,\cdots,y_m)$,则命名实体标注问题可以表示为:在已知序列 H 的条件下,找出使 $Y=(y_1,\cdots,y_t,\cdots,y_m)$的概率 $P(y_1,\cdots,y_t,\cdots,y_m)$最大的序列$[y_1,\cdots,y_t,\cdots,y_m]$。$P(Y|H)$为线性条件随机场,则在随机变量 H 取值 h 的条件下,随机变量 Y 取值为 y 的概率为:

$$P(y|h)=\frac{1}{Z(h)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{t-1},y_i,h,i)+\sum_{i,l}\mu_l s_l(y_i,h,i)\right) \tag{10}$$

$$Z(h)=\sum_y \exp\left(\sum_{i,k}\lambda_k t_k(y_{t-1},y_t,h,i)+\sum_{i,l}\mu_l s_l(y_i,h,i)\right) \tag{11}$$

其中,t_k 和 s_l 是特征函数;λ_k 和 μ_l 是特征值;$Z(h)$为规范化因子,是对所有可能的输出序列求和。

3.4 实体关系抽取(RE)模块

在进行实体的语义关系抽取时,融合实体和实体之间的子序列编码信息,然后将编码特征输入 CNN 模型。该过程可以表示为:

$$R=CNN([e^i,h_{e^i+1},\cdots,h_{e^i+k},\cdots,h_{e^j-1},e^j]) \tag{12}$$

其中,R 表示两个实体之间对应的语义关系,$R\in$ {RES-ATT, ATT-PRA,RES-VAL,N};e^i 和 e^j 表示实体的编码信息;h_{e^i+k}表示词的 Bi-LSTM 的编码特征。一个实体由多个词构成时,用每个词的 Bi-LSTM 特征求和来表示一个实体特征,表示为:

$$e=\sum_{i=B}^{E} h_i \tag{13}$$

其中,B 表示实体开始词的位置,E 表示实体结束词的位置,h_i 表示词 w_i 的编码特征。

CNN 结构如图 4 所示,其中 $w_c^{(i)}\in R^{k\times d}$ 表示第 i 个卷积核,k 表示卷积核的上下文窗口大小,$b^{(i)}$ 表示第 i 个卷积层的偏置单元。因此,对输入特征序列 $S=[e^i,h_{e^i+1},\cdots,h_{e^i+k},\cdots,h_{e^j-1},e^j]$进行卷积操作后,获得潜在特征 $z^{(i)}$。卷积过程表示为:

$$z^{(i)}=\sigma(w_c^{(i)}*s_{l:l+k-1}+b^{(i)}) \tag{14}$$

其中,$z^{(i)}$ 表示通过卷积核 $w_c^{(i)}$ 抽取到的词 s_l 到词 $s_{l:l+k-1}$ 的特征。输入语料的潜在特征可以表示为 $z^{(i)}=[z_1^{(i)},\cdots,z_{L-k+1}^{(i)}]$。在完成卷积操作后,紧接着进行池化(max-pooling)操作来提取经过 $w_c^{(i)}$ 卷积操作的主要特征。池化过程表示为:

$$z_{\max}^{(i)}=\max\{z_1^{(i)},\cdots,z_{L-k+1}^{(i)}\} \tag{15}$$

该模型用多重卷积来抽取多重特征,因此一个输入序列的关系特征可以进一步表示为:

$$R_s=[z_{\max}^1,\cdots,z_{\max}^{(nr)}] \tag{16}$$

其中,nr 表示卷积数目。

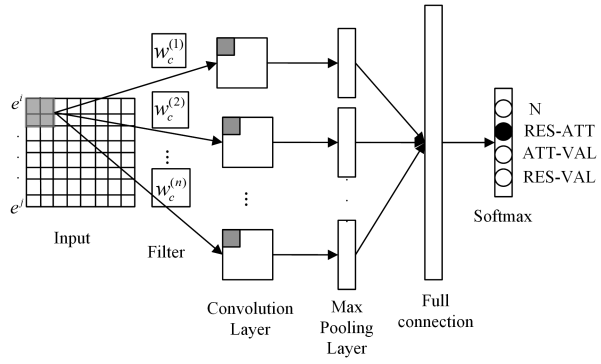


图 4 关系抽取卷积模型

Fig. 4 Convolutional module for relation extraction

最后一层为 softmax,该层将依据关系特征 R_s 计算出该实体对映射到每一种类别上对应的概率。该过程可表示为:

$$y_r=w_R\cdot(R_s\circ r)+b_R \tag{17}$$

$$p_r^i=\frac{\exp(y_r^i)}{\sum_{j=1}^{nc}\exp(y_r^j)} \tag{18}$$

其中,w_R 是 $softmax$ 矩阵,nc 是分类的总数目。

4 算法应用与结果分析

4.1 数据集的建立

本文的实验数据来自于百度百科和从专利文本中爬取到的语料,共有 26399 句资源描述文本,涉及化学药品、材料、电器元件,数据分布如表 3 所列。其中共有实体 96072 个,包括资源实体 12864 个,属性实体 38620 个,参数实体 44588 个。实验采用交叉验证的方法,随机选择 11480 句语料作为训练集,1635 句语料作为验证集,3284 句语料作为测试集,它们各占总语料的 70%,10%和 20%。对于实体关系抽取语料,以同样的比例划分数据集。

表 3 语料领域

Table 3 Corpus domain

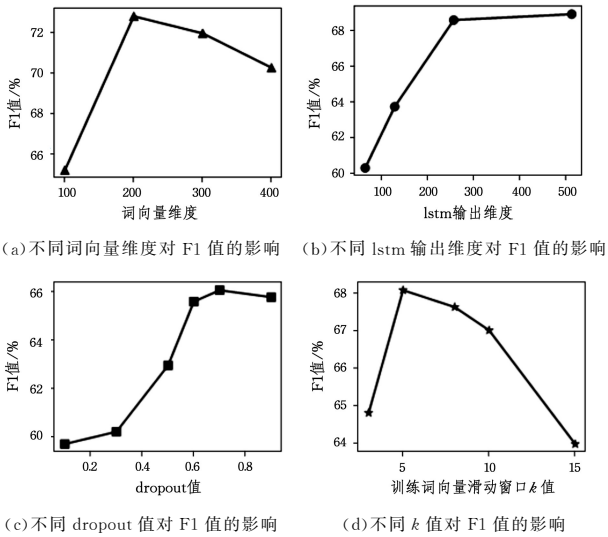
领域	比例/%
化学品	57
材料	25
电器元件	12
其他	6

4.2 实验环境与参数

本文的实验环境如下:算法使用 python3.6 实现;深度学

习库 tensorflow1.4，keras2.1；GPU GTX1080，CPU core i7 四核；内存 16GB。

首先使用 jieba 分词工具包对原始文本做分词处理，然后用 python 的 gensim 工具包对分词结果进行词向量训练，选择 Word2vec 的 Skip-Gram 模型。Att-Bi-LSTM 和 CNN 模型用 keras 实现，并使用反向传播算法进行训练。依据图 5 所示的实验结果设置模型参数，如表 4 所列。



(a)不同词向量维度对 F1 值的影响　(b)不同 lstm 输出维度对 F1 值的影响

(c)不同 dropout 值对 F1 值的影响　(d)不同 k 值对 F1 值的影响

图 5　不同参数值对 F1 值的影响

Fig. 5　Influence of different parameter values on F1 value

表 4　联合抽取模型的参数

Table 4　Parameters of joint extraction model

参数	参数描述	参数值
n_emb	词向量维度	200
k	训练词向量滑动窗口大小	5
n_input	输入对齐 padding 大小	200
drop	Droutout 值	0.7
batch	批处理大小	32
n_lstm	Lstm 单元输出维度	256
n_filter	卷积核数目	128
s_filter	卷积核大小	[3,4,5]
epochs	训练次数	5000

4.3　实验结果分析

管道方式实体语义关系抽取的性能依赖于实体识别的效果。图 6 中的 8 组实验结果证明，实体识别错误会传递到实体语义关系抽取，实体识别错误率越低，抽取到的实体语义关系越准确，因此实体关系抽取结果的好坏可以作为实体识别效果的一个评价指标。实体识别作为实体关系抽取的前提，实体关系抽取的结果又可以反馈调节实体识别模型。

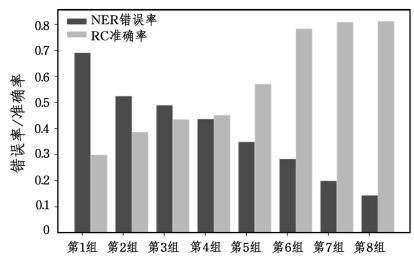


图 6　NER 模型错误率对 RE 的影响

Fig. 6　Influence of NER model accuracy on RE

本文采用精确率 P、召回率 R、F1 值作为评价标准。在探究本文所提实体和关系联合抽取模型 MFB- JoinModel 的性能的过程中，将 Dan[21] 提到的 Pipeline 模型和 Miwa[5] 提到的 SP-Tree 方法作为基线进行对比实验。Pipeline 模型用到的方法是通过训练线性条件随机场模型做命名实体识别，用最大熵模型做关系抽取，先进行命名实体识别，实体识别完成后再进行实体关系抽取。SP-Tree 模型用两个实体词之间的子序列，依据句法结构构成的树形 LSTM 网络来做关系抽取。Join 模型移除了 NER 和 RE 模块的反馈机制，只共享底层提取的特征。在测试集上的实验结果如表 5 所列。

表 5　联合抽取的实验结果

Table 5　Results of combined extraction experiments

（单位：%）

模型	精确率	召回率	F1 值
Pipeline	78.41	63.40	70.11
SP-Tree	81.12	64.67	71.96
Join	79.93	64.05	71.11
MBF-JoinModel	81.03	**68.13**	**74.02**

由实验结果可知，本文提到的 MBF-JoinModel 与已有的联合抽取方法相比，在保证精确率大致相同的前提下可以有效地提升 F1 值，F1 值为 74.02%，相比于管道方式提升了 3.91%；与依据句法结构的树形 LSTM 模型相比，尽管准确率略有衰减，但是召回率和 F1 值都有提高，F1 值可提升 2.06%；相比于 Join 模型，引入反馈机制后，F1 值提升了 2.91%。其中，反馈机制可以有效提升识别模型的 F1 值。从实验结果还可以看出，基于神经网络的模型的性能优于完全基于特征的管道模型的性能，因此用循环神经网络的方法来提取句子特征可以有效地解决实体和关系的联合抽取问题。

图 7 中，管道模型的 NER 模块在 250 min 时趋于收敛，RE 模块在 300 min 时开始趋于收敛，因为两个模块独立运行，所以总体耗时 550 min；联合模型在 400min 时趋于收敛，在总耗时方面联合模型的效率更高。

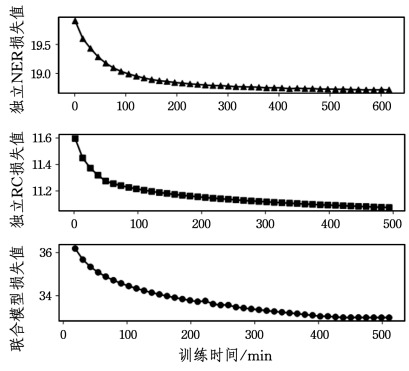


图 7　管道模型和联合模型的收敛时间

Fig. 7　Convergence time of pipe model and joint model

表 6 和表 7 是对联合抽取中实体识别模块和关系抽取模块的独立实验统计。从实验结果看，对属性和量值实体的识别效果整体好于对资源实体的识别效果。这是因为，化工资源名种类繁多，包括氧化物、聚合物、俗称等，形式多样；而描述资源实体的属性和量值实体格式、描述方式相对统一，在大样本环境下相对容易识别。深度学习方法的识别效果优于完

全基于人工特征的 CRF 方法的效果，从而说明了 CRF 方法存在特征表示不充分的问题。神经网络的方法可以在无人工干预的情况下有效地考虑每一个词的上下文信息来提取词特征。Bi-LSTM-CRF 的识别效果优于 RNN-CRF 的识别效果，说明在句子较长即时间步过长时，RNN 模型不能很好地保存长距离依赖性。LSTM 单元通过引入输入门、输出门和遗忘门的概念，强化了长期记忆功能，弥补了 RNN 的缺点，表现出了更好的识别性能。本文提到的联合抽取模型 Mufeedback-JoinModel 引入了 Attention 机制，其目的仍然是解决长距离所带来的问题。Attention 机制可以在输出结果时专注考虑输入序列中的一些被认为是比较重要的词。实验结果证明本文方法能取得比其他方法更好的效果。

表 6 实体识别结果

Table 6 Entity recognition results

（单位：%）

实验	实体类型	精确率	召回率	F1 值
CRF	资源	73.53	70.32	71.89
	属性	80.9	72.07	76.23
	量值	74.19	73.33	73.76
RNN-CRF	资源	79.81	70.83	75.05
	属性	88.61	77.73	82.81
	量值	85.21	78.49	81.72
Bi-LSTM-CRF	资源	82.54	76.36	79.33
	属性	**90.79**	77.66	83.71
	量值	**87.64**	79.91	83.6
MF-Join	资源	**84.11**	**78.47**	**81.19**
	属性	88.12	**80.47**	**84.12**
	量值	87.36	**83.41**	**85.34**

表 7 关系抽取实验结果

Table 7 Experimental results of relation extraction

（单位：%）

模型	精确率	召回率	F1 值
F-CNN-h	73.22	62.18	67.25
S-CNN-h	**75.86**	70.43	73.04
S-CNN-w	75.02	69.44	72.12

表 7 列出了对关系抽取的实验结果，实验分为 3 组，分别选用不同的输入特征。F-CNN-h 使用经过 ATT-Bi-LSTM 编码后的整个句子作为输入特征；S-CNN-h 使用编码后的两个实体之间的子序列作为输入特征；S-CNN-w 模型中，实体词采用 ATT-Bi-LSTM 编码后的特征作为 CNN 的输入，实体对之间的非实体词采用词向量特征作为 CNN 的输入，且每一组实验都融入了实体识别结果和词性作为辅助特征。由实验结果可知，使用两个实体词以及实体词之间的词构成的子序列作为关系抽取模型的输入时的性能优于整个语料作为输入时的性能，这是因为一个句子中可能会包含多个实体，在抽取实体对$\langle e_i, e_j \rangle$的关系时，无关实体 e_k 的特征会对分类产生较大影响，因此尽可能地去除干扰因素，只保留两个实体及实体之间的子序列作为输入。尽管舍掉了实体两端的序列，但是实体的特征仍然经过 ATT-Bi-LSTM 编码层得到，因此子序列仍然包含着实体词的上下文依赖信息。第三组实验中 S-CNN-w 的性能不如其在第二组中的性能，由此说明仅仅使用词向量特征会存在特征表示不充分的问题，仍需要特征提取层来提取上下文特征。

接下来进一步探究实体距离对 3 种关系（RES-ATT，ATT-VAL，RES-VAL）抽取 F1 值的影响，实验结果如图 8 所示。

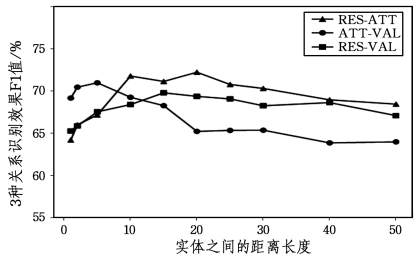


图 8 实体距离对实体关系抽取的影响

Fig. 8 Influence of entity distance on entity relation extraction

从图 8 中可知，实体距离对 ATT-VAL 关系的识别效果的影响较大，这是因为句子中存在属性和量值空间距离比较紧密而其他两种关系空间位置相对稀疏的问题。因此，可以在对 ATT-VAL 关系识别的过程中设置阈值 L_{max}，当实体距离大于 L_{max} 时，认为这两个实体不存在 ATT-VAL 关系。由实验结果图可设置 $L_{max}=15$，当大于这个值时，识别效果显著降低。

结束语 该文提出了基于深度学习的引入反馈机制的混合神经网络模型来实现资源实体及关系的联合抽取。相比于以往基于规则的和基于统计的管道式的模型以及无反馈机制的联合模型，该模型具有无需任何手动特征、减少错误传播概率和有效提取长距离实体之间关系的优点。基于百科爬取到的文本和专利数据集的实验证明，本文所提方法是有效的，可以有效提取实体和实体关系。

但是，联合模型参数较多，对计算设备的要求较高；同时，在实验中可以看出句法结构也是一个重要特征。未来，我们将继续研究如何更好地联合 NER 和 RE 模型，融入句法信息，改进关系抽取的实体表示方法，考虑多实体对算法的影响，以进一步提高实验结果的精确率和 F1 值。

参 考 文 献

[1] SUNDHIM B M. Named Entity task definition, version 2.1 [C]// Proc. of the Sixth Message Understanding Conf. America: Morgan Kaufmann Publishers, 1995: 319-332.

[2] ZHANG L, ZHAO H. Named entity recognition for Chinese microblog with convolutional neural network [C]// International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. China: IEEE Press, 2017: 87-92.

[3] CHEN Y, ZHENG D Q, ZHAO T J. Chinese relation extraction based on Deep Belief Nets[J]. Journal of Software, 2012, 23(10): 2572-2585. (in Chinese)
陈宇，郑德权，赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报，2012，23(10)：2572-2585.

[4] MIWA M, SASAKI Y. Modeling Joint Entity and Relation Extraction with Table Representation [C]// Conference on Empirical Methods in Natural Language Processing. Qatar: Association for Computational Linguistics, 2014: 944-948.

[5] MIWA M, BANSAL M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures [C]// Meeting of the

Association for Computational Linguistics. Germany：dblp：computer science bibliography，2016：1105-1116.

[6] ZHENG S，HAO Y，LU D，et al. Joint Entity and Relation Extraction Based on A Hybrid Neural Network[J]. Neurocomputing，2017，257：1-8.

[7] BAHDANAU D，CHO K，BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J/OL]. Computer Science，2014. https：//arxiv. org/pdf/1409. 0473v6. pdf.

[8] KAMBHATLA N. Combining lexical，syntactic，and semantic features with maximum entropy models for extracting relations [C]// ACL Interactive Poster Demonstration Sessions，Spain：Association for Computational Linguistics. 2004：22-25.

[9] OUDAH M，SHAALAN K. NERA 2. 0：Improving coverage and performance of rule-based named entity recognition for Arabic[J]. Natural Language Engineering，2017，23：441-472.

[10] ZHOU G D，SU J. Named entity recognition using an HMM-based chunk tagger[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. America，2002：473-480.

[11] MCCALLUM A，LI W. Early results for named entity recognition with conditional random fields，feature induction and web-enhanced lexicons[C]// Conference on Natural Language Learning at Hlt-Naacl. Canada：Association for Computational Linguistics，2003：188-191.

[12] ZHANG H N，WU D Y，LIU Y，et al. Chinese Named Entity Recognition Based on Deep Neural Network[J]. Journal of Chinese Information Processing，2017，31(4)：28-35. (in Chinese)
张海楠，伍大勇，刘悦，等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报，2017，31(4)：28-35.

[13] DONG C H，ZHANG J J，ZONG C Q. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[C]// International Conference on Computer Processing of Oriental Languages. Lecture Notes in Computer Science，Cham：Springer，2016：239-250.

[14] LUO L，YANG Z，YANG P，et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics，2017，34(8)：1381-1388.

[15] CHE W X，LIU T，LI S. Automatic Entity Relation Extraction [J]. Journal of Chinese Information Processing，2005，19(2)：1-6. (in Chinese)
车万翔，刘挺，李生. 实体关系自动抽取[J]. 中文信息学报，2005，19(2)：1-6.

[16] MA X J，GUO J Y，XIAN Y T，et al. Entity Hyponymy Acquisition and Organization Combining Word Embedding and Bootstrapping in Special Domain[J]. Computer Science，2018，45(1)：67-72. (in Chinese)
马晓军，郭剑毅，线岩团，等. 结合词向量和 Bootstrapping 的领域实体上下位关系获取与组织[J]. 计算机科学，2018，45(1)：67-72.

[17] SOCHER R，PENNINGTON J，HUANG E H，et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]// Conference on Empirical Methods in Natural Language Processing(EMNLP 2011). UK：DBLP，2011：151-161.

[18] LAI S，XU L，LIU K，et al. Recurrent convolutional neural networks for text classification[C]// Conference of the Association for the Advancement of Artificial Intelligence (AAAI). America：DBLP，2015：2267-2273.

[19] YAN X，MOU L，LI G，et al. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Path [J]. Computer Science，2015，42(1)：56-61.

[20] LI J W，LUONG M T，JURAFSKY D，et al. When Are Tree Structures Necessary for Deep Learning of Representations? [J/OL]. https：//arxiv. org/abs/1503. 00185.

[21] DAN R，YIH W T. 1 Global Inference for Entity and Relation Identification via a Linear Programming Formulation[M]// Introduction to Statistical Relational Learning. MIT Press，2007：608-636.

[22] YANG B，CARDIE C. Joint Inference for Fine-grained Opinion Extraction[C]// Meeting of the Association for Computational Linguistics. Bulgaria：ACL，2013：1640-1649.

[23] SINGH S，SINGH M，CHANANA S，et al. Operation and control of a hybrid wind-diesel-battery energy system connected to micro-grid[C]// International Conference on Control，Automation，Robotics and Embedded Systems. India：IEEE Press，2013：1-6.

[24] LI Q，JI H. Incremental Joint Extraction of Entity Mentions and Relations[C]// Meeting of the Association for Computational Linguistics. America：BibSonomy，2014：402-412.

[25] DZMITRY B，KYUNGHYUN C，YOSHUA B，et al. Neural machine translation by jointly learning to align and translate [J/OL]. https：//arxiv. org/abs/1409. 0473v2.