一种基于混合神经网络的命名实体识别与共指消解联合模型

郜成胜[1],张君福[1],李伟平[1],赵 文[2],张世琨[2]
(1. 北京大学软件与微电子学院,北京 100871; 2. 北京大学软件工程国家工程研究中心,北京 100871)

摘 要: 命名实体识别与共指消解均依赖于对实体相邻文本信息的学习,本文提出一种基于混合神经网络的命名实体识别与共指消解联合模型,共用双向长短时记忆模型 LSTM 编码层对输入序列中每个词前后方向上下文信息进行编码,并通过训练学习得到上下文信息传递到前馈神经网络 FFNN 模型以提高共指消解精度,通过将领域文档及篇章语义向量加入 FFNN,改进共指消解算法并优化共指消解模型. 基于领域文本数据集进行联合模型训练,实验结果表明该联合模型可以有效地提高共指消解精度.

关键词: 神经网络; 命名实体识别; 共指消解; 联合神经网络模型
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2020)03-0442-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.03.004

A Joint Model of Named Entity Recognition and Coreference Resolution Based on Hybrid Neural Network

GAO Cheng-sheng[1], ZHANG Jun-fu[1], LI Wei-ping[1], ZHAO Wen[2], ZHANG Shi-kun[2]
(1. *School of Software and Microelectronics, Peking University, Beijing* 100871, *China*;
2. *National Engineering Research Center for Software Engineering, Peking University, Beijing* 100871, *China*)

Abstract: Considering that both named entity recognition and coreference resolution depend on the same context of the entity word, this paper proposes a hybrid neural network model to settle these problems which contains a named entity recognition (NER) module and a coreference resolution (CR) module. NER and CR share the same bidirectional LSTM encoding layer, which is used to encode each input word by taking into account the context on both sides of the word. The contextual information of entities obtained in BiLSTM encoding layer further pass through to FFNN module to improve the coreference resolution. Furthermore, by adding domain documents and chapter semantic vectors to FFNN, the coreference resolution algorithm is improved and the coreference resolution model is optimized. Finally, we conduct experiments on the domain dataset to verify the effectiveness of our method. The joint model can effectively improve the accuracy of coreference resolution task.

Key words: neural network; named entity recognition; coreference resolution; hybrid neural network model

1 引言

命名实体识别和共指消解是知识抽取研究的重要问题,在自然语言处理和知识图谱构建中起着重要作用. 命名实体识别是指从文本数据集中自动识别出命名实体指称项的过程,共指消解实现将抽取出的多个指称项对应于同一个实体对象. 二者既可独立,也存在关联关系.

采用联合模型解决命名实体识别和共指消解成为一种提高共指消解能力的新研究思路,联合模型可以实现命名实体识别和共指消解有效关联,充分利用命名实体识别过程信息提高共指消解能力[1]. Pascal Denis 等[2]提出了采用 ILP 方法对共指消解和实体分类建立联合模型作为全局推理问题进行研究; Gang Luo 等[3]提出命名实体识别和链接联合方法,获取命名实体识别和实体链接之间相互依赖性,利用一项任务信息提

收稿日期:2018-09-24;修回日期:2018-11-22;责任编辑:梅志强

高另一项任务能力. 现有联合模型大多基于特征或者知识库进行命名实体识别和链接,需要依靠复杂的人工特征建模工程,易引起错误蔓延. 为减少对人工建模的依赖,相关研究在 POS、NER 等序列标记任务中提出了采用 RNN、Long Short-Term Memory (LSTM)等模型学习自然语言文本上下文信息的深层网络模型方法[3];在共指消解任务研究中提出了端到端深层网络模型[4].

基于以上分析及实际工程需求,本文提出一种基于混合神经网络的命名实体识别与共指消解联合模型,命名实体识别和共指消解共用一个双向 LSTM 编码层,该编码层对输入序列中每个词的前后方向上下文信息进行编码. 命名实体识别的解码层对每个实体标签进行独立预测,双向 LSTM 将命名实体识别编码层训练学习得到的上下文信息传递到共指消解解码层,共指消解解码层通过 FFNN 网络模型生成 mention-pair 分布式表示及其共指评分,通过全部共指评分条件概率分布获取最优的共指关系进行共指消解. 本文提出的联合模型基于领域自然语言文本数据集进行了训练和测试,实验结果证实该联合模型可以有效提升共指消解能力.

2　相关工作

命名实体识别和共指消解是自然语言文本知识库和知识图谱构建研究中的重要内容,有助于自然语言处理有关应用的研究,存在两种不同研究架构:流水线方法和联合学习方法. 流水线方法将命名实体识别和共指消解作为两个独立的任务顺序进行,过去几年相关研究较多采用该方法,例如命名实体识别[5~8]和共指消解[9~12].

命名实体识别方法主要包括基于规则方法和基于统计机器学习方法,已有研究中采用的统计模型主要包括隐马尔科夫模型(HMM)、条件马尔科夫模型(CMM)和条件随机场(CRFs)等[13~15],此类方法抽取效果非常依赖于人工特征建模工作和外部知识库质量. 近来开始将多种神经网络应用于命名实体识别研究,将命名实体识别任务视为序列标记任务. Collobert 等在词表示基础上使用 CNN 和 CRF 进行命名实体识别工作[16];Chiu 等[17]提出了一种 LSTM 和 CNN 联合网络实现命名实体识别,采用 CNN 完成构成词的字符表示学习,采用双向 LSTM 学习每个词前后向上下文信息,采用一个线性输出层和一个逻辑函数层进行实体标签推测;Huang 等[3]提出采用双向 LSTM 进行编码和 CRF 解码的命名实体识别模型,取得了较好效果. 以上模型均采用双向 LSTM 作为编码模型,解码方式采用不同网络模型.

基于机器学习的共指消解研究已经有较长历史[18],共指消解的特征学习问题一直是研究重点,先后出现了自动生成解析树结合人工建模方法[19]、词法特征学习[20]等方法,近年来神经网络模型应用[21]取得了较好的效果. 已有研究中大多集中于 mention 对之间关系学习,学习目标主要限定于 mention 特征,最新研究主要包括扩展学习文本范围和文本特征结合等方向. 文献[4]探索了基于 span embedding 端到端深层网络共指消解模型;文献[21]提出在 encoding 层之上结合相关特征信息建立 mention-pair 分布式表示及其共指评分,通过全部共指评分的条件概率分布获取最优的共指关系,提高共指消解能力.

联合模型的研究在统计机器学习时代已经出现,文献[22]基于 Dynamic CRFs 建立联合模型用于同时进行词性标注和词块分析等多项序列标记任务,支持多项任务间共享信息和近似推理,取得了较线性链 CRFs 更高的词性标注精度;Finkel 等[23]提出了语法解析和实体标注联合模型,解决了两项任务的一致性问题,同时联合模型可以支持使用一项任务信息提高其他任务表现;Sil 等[24]研究了命名实体识别和链接联合模型,命名实体识别方法与 Freebase 知识库结合,将实体间链接依赖性与实体边界联合判断,取得了较好结果. 文献[25]提出 LaSO 模型,将全局特征融入基于检索的共指消解预测方法,探索了命名实体识别与共指消解联合模型方法,降低了实体解析错误的级联扩散问题.

3　联合模型

为了有效提高共指消解精度,本文提出一种基于混合神经网络的命名实体识别与共指消解联合模型,如图 1 所示. 联合网络模型的第一层是命名实体识别和共指消解共用的基于双向 LSTM 的编码层,之上包括两个输出通道将编码层分别连接到命名实体识别模块和共指消解模块. 命名实体识别模块采用 LSTM 网络进行解码,共指消解模块通过 Feed Forward Neural Network (FFNN)网络实现.

3.1　基于双向 LSTM 网络编码层

编码层由词表示层、forward LSTM 层、backward LSTM 层和连接层组成. 词表示层生成词向量表示,词序列可表示为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\cdots,\boldsymbol{x}_n\}$,$\boldsymbol{x}_t\in\mathbb{R}^d$ 为句子中第 t 个词的 d 维词向量表示. 在词表示层之上是两个平行的 LSTM 层:forward LSTM layer 和 backward LSTM layer. 对应每个词 $\boldsymbol{x}_t$,考虑词的上下文,forward LSTM layer 从 $\boldsymbol{x}_1$ 到 $\boldsymbol{x}_n$ 对 $\boldsymbol{x}_t$ 进行编码,标记为 $\overleftarrow{\boldsymbol{h}}_t$,backward LSTM layer 从 $\boldsymbol{x}_n$ 到 $\boldsymbol{x}_1$ 对 $\boldsymbol{x}_t$ 进行编码,标记为 $\overrightarrow{\boldsymbol{h}}_t$.

LSTM 由一系列循环连接的子网络组成,这些子网络称为存储块,forward 隐层和 backward 隐层中的每个时刻的状态都可以表示为一个存储块. 每个存储块包

括一个或者多个自连接的判断信息有用与否的存储单元和3个门控单元提供对存储单元进行写入、读取和重置类似的操作,3个门控单元包括输入门、输出门和遗忘门. 在每个时刻的LSTM存储块都基于前一个时刻隐层向量$\boldsymbol{h}_{t-1}$、前一个时刻存储单元向量$\boldsymbol{c}_{t-1}$和当前时刻输入的词表示$\boldsymbol{x}_t$,分别可以声明为$\overrightarrow{\boldsymbol{h}}_t=\text{lstm}(\overrightarrow{\boldsymbol{h}}_{t-1},\overrightarrow{\boldsymbol{c}}_{t-1},\boldsymbol{x}_t)$和$\overleftarrow{\boldsymbol{h}}_t=\text{lstm}(\overleftarrow{\boldsymbol{h}}_{t+1},\overleftarrow{\boldsymbol{c}}_{t+1},\boldsymbol{x}_t)$. lstm功能详细实现如下所示:

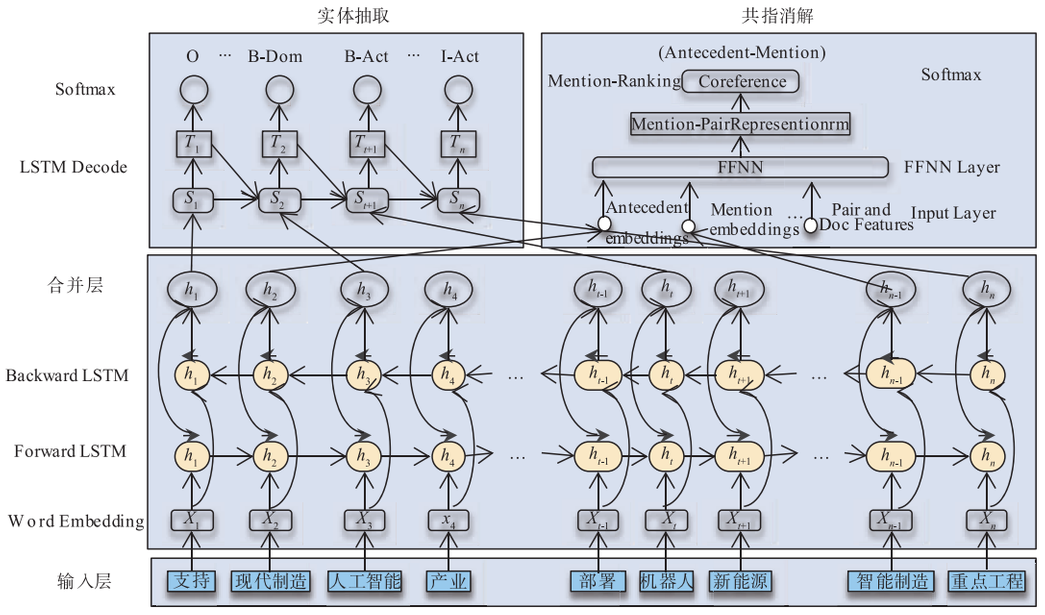

图1 命名实体识别与共指消解联合模型

$$
\begin{aligned}
i_t&=\sigma(w_{xi}x_t+w_{hi}h_{t-1}+w_{ci}c_{t-1}+b_i)\\
f_t&=\sigma(w_{xf}x_t+w_{hf}h_{t-1}+w_{cf}c_{t-1}+b_f)\\
z_t&=\tanh(w_{xc}x_t+w_{hc}h_{t-1}+b_c)\\
c_t&=f_tc_{t-1}+i_tz_t\\
o_t&=\sigma(w_{xo}x_t+w_{ho}h_{t-1}+w_{co}c_t+b_0)\\
h_t&=o_t\tanh(c_t)
\end{aligned}
$$

i、f和O分别表示输入门、遗忘门和输出门,b表示偏置单元,c表示存储单元. σ和tanh分别为sigmoid和tanh激活函数.

将$\overrightarrow{\boldsymbol{h}}_t$和$\overleftarrow{\boldsymbol{h}}_t$拼接后形成第$t$个词的编码表示,可以声明为$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t]$.

3.2 命名实体识别模型

命名实体识别为待识别文本中的每个词分配一个实体标签,实体标签针对实体类别结合编码体系进行定义,每个标签既包含实体类别信息,也包含词在实体中的位置信息,采用LSTM网络直接对每个词的标签分配过程进行建模.

在为一个词t指定实体标签的过程中,命名实体识别解码层的输入信息包括:获取自编码层的合并后的词向量$\boldsymbol{h}_t$、前一时刻的标签预测向量$\boldsymbol{T}_{t-1}$、LSTM模型的前一时刻隐层向量$\boldsymbol{S}_{t-1}$,每个LSTM单元与解码层的LSTM单元基本一致,只有输入门的输入信息有变化,输入门改写如下:

$$i_t=\sigma(\boldsymbol{w}_{hi}h_t+\boldsymbol{w}_{si}s_{t-1}+\boldsymbol{w}_{ti}T_{t-1}+b_i)$$

$\boldsymbol{w}_{hi}$、$\boldsymbol{w}_{si}$、$\boldsymbol{w}_{ti}$为LSTM单元对应输入信息权重变量,σ为sigmoid激活函数.

标记预测向量$\boldsymbol{T}$由隐层状态$\boldsymbol{S}$转换得到:

$$\boldsymbol{T}_t=\boldsymbol{W}_{ts}\boldsymbol{S}_t+b_{ts}$$

最后,在全连接层,依据标签预测向量$\boldsymbol{T}_t$计算实体标签概率:

$$
\begin{aligned}
y_t&=\boldsymbol{W}_y\boldsymbol{T}_t+b_y\\
P_t^i&=\frac{\text{Exp}(y_t^i)}{\sum_{j=1}^{\text{nt}}\text{Exp}(y_t^j)}
\end{aligned}
$$

此处$\boldsymbol{W}_y$为全连接层(softmax层)的权重矩阵,nt是标签总数. 鉴于标签预测向量$\boldsymbol{T}$可以视为标签的embedding,LSTM可学习到标签的远距离依赖性,该模型通过实体标签分配建模,实现命名实体识别过程建模.

3.3 共指消解模型

共指消解模型由mention-pair编码系统和mention-ranking模型组成,mention-pair编码系统通过FFNN网络模型处理相关输入,生成mention-pair的分布式表示,mention-ranking模型处理mention-pair分布式表示,生成mention-pair评分,通过对所有mention-pair评分的条件概率分布获取最优共指关系.

3.3.1 mention-pair 编码系统

mention-pair编码系统通过FFNN网络模型对

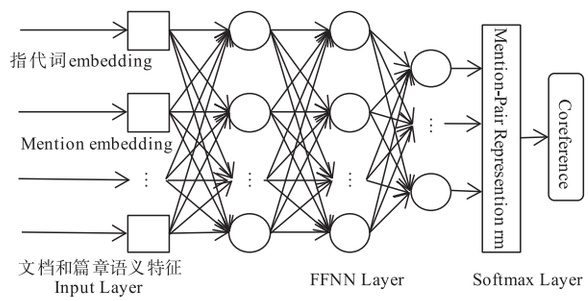

图2　基于FFNN的共指消解模型

mention m 和候选指代词 a 生成 mention-pair 的分布式表示 $\boldsymbol{r}_m(a,m)\in\mathbb{R}^d$. 主要包括输入层和隐层两部分.

(1)输入层

输入层主要将 mention 包含的不同词及多个词信息输入 FFNN 网络,每个词通过一个向量表示,用多个词的向量平均值来表示 mention,另外包括 mention 在句子中的距离、长度等特征以及 mention-pair 的二元特征. 距离和长度特征可分类为[0,1,2,3,4,5-7,8-15,16-31,32-63,64+],可以包含到连续特征中,也可以编码为 1-hot 向量. 考虑共指关系与整个文档及篇章上下文信息相关,本文结合领域实际需求,通过 TFIDF 算法提取文档及篇章主题作为输入特征,提高共指消解精度. 主要包含特征如表 1.

表 1　共指消解相关特征

特征类型	包含内容
embedding 特征	包括 mention 的首词的 word embedding、末尾词的 word embedding、前词的 word embedding、后词的 word embedding
mention 附加特征	mention 词性(代词、名词、专有名词等)、mention 位置信息(文档中的顺序位置)、是否嵌套、mention 中词的长度等
文档信息	包括文档的类型等,例如规划、报告、参考文献等类型
距离特征	mention 在句子中的距离、mentions 之间距离以及 mentions 是否存在重叠等特征信息
词匹配特征	mention 所包含字符的匹配信息
语义特征	领域文本文档语义信息、及篇章语义信息

将以上特征向量全部拼接在一起形成 I 维向量 $\boldsymbol{h}_0$,作为 mention-pair 表示学习神经网络的输入:

$$\boldsymbol{h}_0(a,m)=\mathrm{FFNN}([\boldsymbol{x}_a^*,\boldsymbol{\phi}_a,\boldsymbol{x}_m^*,\boldsymbol{\phi}_m,\boldsymbol{\phi}(a,m)])$$

其中 $\boldsymbol{x}_a^*$ 为指代词 a 的 word embedding,$\boldsymbol{\phi}_a$ 为指代词 a 的特征向量,$\boldsymbol{x}_m^*$ 为 mention m 的 word embedding,$\boldsymbol{\phi}_m$ 为 mention m 的特征向量,$\boldsymbol{\phi}(a,m)$ 为 mention-pair 的特征向量.

(2)隐层

输入层合并形成的输入信息通过 3 层 sigmoid 单元隐层处理,所有隐层单元与前一层之间是全连接结构:

$$\boldsymbol{h}_i(a,m)=\delta(\boldsymbol{W}_i\boldsymbol{h}_{i-1}(a,m)+b_i)$$

$\boldsymbol{W}_1$ 是 $M_1\times I$ 权重矩阵,$\boldsymbol{W}_2$ 是 $M_2\times M_1$ 权重矩阵,$\boldsymbol{W}_3$ 是 $d\times M_2$ 权重矩阵. 最后一个隐层的输出就是 mention-pair 的向量表示:

$$r_m(a,m)=h_3(a,m)$$

(3)篇章主题特征提取

通过与已指定主题的篇章文本进行相似度计算,为领域文本指定篇章主题. 初始篇章主题通过领域文本中具有明确篇章标题包含的特征主题词采用算法处理结合人工验证方法指定. 篇章相似度计算采用 TF-IDF 算法[23]和余弦相似度算法.

词频 tf 指的是特征项在所有文档中出现的频率,计算公式为:

$$\mathrm{tf}_i=\frac{n_i}{\sum_{k=1}^{m}n_k}$$

其中,n_i 是词 t_i 在所有文档和篇章中出现的次数,$\sum_{k=1}^{m}n_k$ 是所有文档和篇章中所有词数量的和.

逆向文件频率 idf 是对一个词的普遍重要性的度量,计算公式为:

$$\mathrm{idf}_i=\log\left(\frac{N}{n_i}+0.01\right)$$

w_{ij} 定义为 TFIDF 因子,计算公式为:

$$w_{ij}=\frac{\mathrm{tf}_i\times\log\left(\frac{N}{n_i}+0.01\right)}{\sqrt{\sum_{i=1}^{N}\mathrm{tf}_i^2\times\log\left(\frac{N}{n_i}+0.01\right)^2}}$$

篇章相似度余弦距离计算公式为:

$$\mathrm{sim}(d_1,d_2)=\frac{\sum_{k=1}^{n}w_{ki}w_{kj}}{\sqrt{\sum_{k=1}^{n}w_{ki}^2}\sqrt{\sum_{k=1}^{n}w_{kj}^2}}$$

$$0\leqslant\mathrm{sim}(d_1,d_2)\leqslant1$$

3.3.2　mention-ranking 模型

mention-ranking 模型为每个 mention 指定一个评分最高的备选实体. mention-ranking 模型为每个 mention m 和一个备选实体 a 生成评分 $s_m(a,m)$ 表示二者之间的共指关系符合度.

$$s_m(a,m)=\boldsymbol{W}_m r_m a,m+b_m$$

此处 $\boldsymbol{W}_m$ 是一个 $1\times d$ 的权重矩阵.

通过训练 mention-ranking 模型,将每个 mention 与其共指符合度评分最高的指代实体建立连接.

D 表示要处理的目标文档,其中包含的 mention 数量为 N,为每个 mention m 指定一个指代词 y_m,则 y_m 的所有指派方式集合为 $Y(m)=\{\varepsilon,1,\cdots,m-1\}$,$\varepsilon$ 表示

虚拟的指代词,适用于当 mention 不是一个实体或者 mention 不存在指代词的情况.

考虑 m 为文档中第一个实体的情况,即 m 之前无指代实体,mention m 和备选实体 a 的符合度评分 $s(a,m)$ 可以表示为:

$$s(a,m)=\begin{cases}0, & a=\varepsilon\\ s_m(a,m), & a\neq\varepsilon\end{cases}$$

通过条件概率分布的最优配置得到实体与指代实体共指关系概率 P_{CR},用如下多项式表示:

$$P_{\mathrm{CR}}=P(y_1,\cdots,y_N\mid D)=\prod_{i=1}^{N}P(y_i\mid D)=\prod_{i=1}^{N}\frac{\mathrm{Exp}(s(i,y_i))}{\sum_{y'Y(i)}\mathrm{Exp}(s(i,y'))}$$

其中 $s(i,y_i)$ 表示 mention i 与指代词 y_i 的共指消解评分,即 $s(a,m)$.

4 训练方法与实验结果

4.1 训练方法

模型训练目标是使联合模型在训练集上达到对数似然函数最大化. 在命名实体识别训练中,采用 RMSprop[26] 作为目标函数.

$$L_{\mathrm{ner}}=\max\sum_{j=1}^{|D|}\sum_{t=1}^{L_j}\log(P_t^{(j)}=y_t^{(j)}\mid X_j,\Theta_{\mathrm{ner}})$$

其中 $|D|$ 是数据集的大小,L_j 是句子 X_j 的长度,$y_t^{(j)}$ 是句子中的词 t 的标签,$P_t^{(j)}$ 是实体标签概率.

在共指消解训练中,采用如下目标函数:

$$L_{\mathrm{cr}}=\max\sum_{j=1}^{|D|}\log(P_{\mathrm{cr}}^{(j)}=y_{\mathrm{cr}}^{(j)}\mid X_j,\Theta_{\mathrm{cr}})$$

$P_{\mathrm{cr}}^{(j)}$ 为实体与指代实体的共指关系概率.

4.2 实验设置

实验基于我国规划领域文本数据集开展,测试指标主要包括准确率、召回率和 $F1$ 测度值,此外实验了深层神经网络不同超参数设置的实验结果.

(1)数据集

本文采用自主采集整理的规划领域文本库作为实验数据集,包括我国十二五、十三五国家、省、市县 3 级总共约 12000 部规划文本,约 2.1 亿字;经过文本预处理、分词处理、去停用词后,形成领域词库,进行训练集和测试集标注. 按照领域命名实体类别进行实体标注,主要包括地域(region)、领域(domain)、战略(strategy)、设施(facility)、指标(index)、工程(project)、机构(organization)等 8 类,采用 BIO 编码方式标注. 在完成命名实体识别后,采用实体对标注的方式进行共指关系标注. 训练集数据由 100 部规划组成,约 500 万字,包括命名实体约 5 万个,共指关系约 1 万个;测试数据由 30 部规划组成,约 150 万字,包括命名实体约 1 万个,共指关系约 4000 个.

(2)指标

采用全匹配正确计数的方法. 只有当一个实体的识别结果与测试集标记完全相同,才记为正确识别,其余情况皆认为错误识别. 采用准确率(Precision,P)、召回率(Recall,R)以及 $F1$ 测度值($F1$-score,$F1$)对命名实体识别和共指消解进行测试统计.

(3)超参数

联合模型使用的超参数包括 word embedding 维数、encoding 层隐层数量、decoding 层隐层数量、FFNN 网络层数和激活函数等. 超参数信息见表 2.

表 2 超参数表

参数	参数描述	参数值
d	word embedding 向量维数	200
nend	encoding、decoding 层隐层数量	200,100
nf	FFNN 网络层数	3
Acti-Fun	FFNN 网络激活函数	sigmoid

4.3 实验结果及分析

实验中,我们对不同超参数分别进行实验优化. 超参数初值设置如下:激活函数(sigmoid),word embedding 向量维数(50),隐层数量([100,50])([encoding 隐层数量,decoding 隐层数量]),FFNN 网络层数(5). 实验结果如表 3 ~ 6 所示.

表 3 word embedding 向量维数超参数实验结果对比

超参数	取值	命名实体识别			共指消解		
		准确率 P	召回率 R	$F1$	准确率 P	召回率 R	$F1$
word embedding 向量维数	50	52.83	46.41	49.41	48.72	43.99	46.23
	100	68.31	59.08	63.36	59.45	50.43	54.57
	200	79.92	68.46	73.75	64.1	59.7	61.82

表 4 隐层数量超参数实验结果对比

超参数	取值	命名实体识别			共指消解		
		准确率 P	召回率 R	$F1$	准确率 P	召回率 R	$F1$
隐层数量[encoding,decoding]	[100,50]	63.58	49.81	55.86	56.92	51.23	53.93
	[100,100]	68.35	52.43	59.34	62.31	56.63	59.33
	[200,100]	79.92	68.46	73.75	64.1	59.7	61.82
	[400,200]	73.35	62.92	67.74	63.07	54.63	58.55

表 5　FFNN 网络层数超参数实验结果对比

超参数	取值	命名实体识别			共指消解		
		准确率 P	召回率 R	$F1$	准确率 P	召回率 R	$F1$
FFNN 网络层数	2	73.71	63.16	68.03	58.12	51.83	54.80
	3	79.92	68.46	73.75	64.1	59.7	61.82
	5	78.65	66.91	72.31	62.41	56.48	59.30

表 6　激活函数超参数实验结果对比

超参数	取值	命名实体识别			共指消解		
		准确率 P	召回率 R	$F1$	准确率 P	召回率 R	$F1$
激活函数	tanh	69.53	60.28	64.58	53.46	48.2	50.69
	sigmoid	79.92	68.46	73.75	64.1	59.7	61.82
	ReLU	71.84	61.73	66.40	55.61	50.83	53.11

结果显示:激活函数为 sigmoid,word embedding 向量维数为 200,隐层数量为[200,100],FFNN 网络层数为 3 时,实验结果最优.

实验主要对比了命名实体识别和共指消解的几类最新方法,包括 HMM[13]、BiLSTM + CRF[3]、Mention-Pair[21]、Span-Score[4]和 LaSO 模型[25]. 前 4 类方法分别将统计方法和神经网络方法应用于命名实体识别和共指消解任务,LaSO 模型为命名实体识别和共指消解联合模型. 上述方法基于领域数据集的实验结果如表 7 所示.

表 7　不同方法实验结果对比

方法	命名实体识别			共指消解		
	准确率 P	召回率 R	$F1$	准确率 P	召回率 R	$F1$
HMM	73.56	62.41	67.53	–	–	–
BiLSTM + CRF	81.40	74.70	77.91	–	–	–
Mention-Pair	–	–	–	59.10	50.60	54.52
Span-Score	–	–	–	61.20	56.90	58.97
LaSO	63.62	53.46	58.10	54.10	47.70	50.70
本文联合模型	79.92	68.46	73.75	64.10	59.70	61.82

实验结果中,BiLSTM + CRF 方法在命名实体识别任务实验结果最好,本文联合模型在命名实体识别任务中 $F1$ 为 73.75,与 BiLSTM + CRF 方法表现优于其他方法;在共指消解任务中 $F1$ 为 61.82,优于其他方法. 结果证明了基于深层神经网络的联合模型方法对命名实体识别和共指消解联合任务有效. LaSO 方法用于领域数据集结果稍差,分析原因在于其主要基于特征检索,领域特征建模不足导致实验结果较差. 实验结果也佐证在处理大规模自然语言文本领域数据集且难于进行深入特征建模时,统计机器学习方法和深度学习方法表现优于传统的基于规则方法,在处理存在先后依存关系的多个任务时,联合模型方法表现优于独立模型.

5　结论

本文提出一种基于深层神经网络的联合模型,在不依靠人工特征建模条件下提升了命名实体识别和共指消解能力. 对比其他深层神经网络,本方法考虑了实体标签全局关联信息识别,并且实现了与实体多种特征信息结合. 为证明方法效果,本文基于领域大规模自然语言文本数据集进行了相关实验,实验结果证明该方法在命名实体识别任务中达到了当前最先进方法,在共指消解任务中超过了已有方法.

后续研究中,我们将通过加强数据标注和超参数调优等工作,进一步改善联合模型识别能力,并且在其他领域数据集进行测试验证模型效能.

参考文献

[1] Luo G, Huang X, Lin C Y, et al. Joint entity recognition and disambiguation[A]. Conference on Empirical Methods in Natural Language Processing[C]. Austin: EMNLP, 2016. 879 – 888.

[2] Denis P, Baldridge J. Global joint models for coreference resolution and named entity classification[J]. Procesamiento del Lenguaje Natural, 2009, 42(1): 87 – 96.

[3] Lample G, Ballesteros M, et al. Neural architectures for named entity recognition[A]. Proceedings of NAACL-HLT 2016[C]. San Diego: NAACL-HLT, 2016. 260 – 270.

[4] Lee K, He L, Lewis M, et al. End-to-end neural coreference resolution[A]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing[C]. Copenhagen: EMNLP, 2017. 188 – 197.

[5] Muslea I. Extraction patterns for information extraction tasks: A survey[EB/OL]. http://aaai.org/Papers/Workshops/1999/WS-99-11/WS99-11, 2018-01-01.

[6] Rau L F. Extracting company names from text[A]. Seventh IEEE Conference on Artificial Intelligence Application[C]. Miami: IEEE, 1991. 29 – 32.

[7] Maynard D, Tablan V, Ursu C. Named entity recognition from diverse Text Types[A]. Recent Advances in Natural Language Processing Conference[C]. Borovets: RANLP, 2001.

[8] Liu X, Zhang S, Wei F, et al. Recognizing named entities in

tweets[J]. ACL,2011,1:359 – 367.
[9] Wagstaff K, Cardie C. Clustering with instance-level constraints. [A] Seventeenth International Conference on Machine Learning[C]. Stanford:ICML,2000. 1103 – 1110.
[10] Ng V. Unsupervised models for coreference resolution. [A]. Conference on Empirical Methods in Natural Language Processing[C]. Singapore:EMNLP,2009. 6 – 7.
[11] Haghighi A, Dan K. Coreference resolution in a modular, entity-centered model[A]. Human Language Technologies:The 2010 Conference of the North American Chapter of the Association for Computational Linguistics. [C]. Los Angeles:NAACL-HLT,2010. 385 – 393.
[12] Haghighi A, Dan K. Unsupervised coreference resolution in a nonparametric bayesian model[A]. Meeting of the Association of Computational Linguistics[C]. Prague: ACL,2007. 848 – 855.
[13] Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution[A]. Proceedings of the Eighteenth Conference on Computational Natural Language Learning[C]. Michigan: CoNLL, 2014. 78 – 86.
[14] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields:probabilistic models for segmenting and labeling sequence data[A] Eighteenth International Conference on Machine Learning[C]. Williamstown: ICML, 2001. 282 – 289.
[15] Colmenar J M, Abanades M A, Poza F, et al. On a generalized name entity recognizer based on hidden markov models[A]. International Conference on Intelligent Systems Design and Applications[C]. Cordoba:ISDA,2011. 952 – 958.
[16] Collobert R, Weston J, Karlen M, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12(1):2493 – 2537.
[17] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4(1): 357 – 370.
[18] Ng V. Supervised noun phrase coreference research: the first fifteen years[A]. Meeting of the Association for Computational Linguistics[C]. Uppsala:ACL,2010. 1396 – 1411.
[19] Raghunathan K, Lee H, Rangarajan S, et al. A multi-pass sieve for coreference resolution. [A]. Conference on Empirical Methods in Natural Language Processing[C]. Massachusetts:EMNLP,2010. 492 – 501.
[20] Greg Durrett, Dan Klein. Easy victories and uphill battles in coreference resolution[A]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing[C]. Seattle:EMNLP,2013. 1971 – 1982.
[21] Clark K, Manning C D. Improving coreference resolution by learning entity-level distributed representations[A]. 54th Annual Meeting of the Association for Computational Linguistics[C]. Berlin:ACL,2016. 643 – 653.
[22] Mccallum A, Rohanimanesh K, Sutton C. Dynamic conditional random fields for jointly labeling multiple sequences [A]. NIPS Workshop on Syntax, Semantics, and Statistics [C]. Vancouver:NIPS,2003.
[23] Finkel J R, Manning C D. Joint parsing and named entity recognition[A]. Human Language Technologies:Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings[C]. Boulder. DBLP,2009. 326 – 334.
[24] Sil A, Yates A. Re-ranking for joint named-entity recognition and linking[A]. ACM International Conference on Conference on Information & Knowledge Management [C]. San Francisco:CIKM,2013,2369 – 2374.
[25] Iii H D, Marcu D. A large-scale exploration of effective global features for a joint entity detection and tracking model[A]. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing[C]. Vancouver: HLT/EMNLP,2009. 97 – 104.
[26] Tieleman T, Hinton G. Lecture 6. 5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural Networks For Machine Learning, 2012,4(2):26 – 31.

作者简介

郜成胜 男,1975 年出生,山西长治人,现为北京大学软件与微电子学院工程博士研究生,主要研究领域为知识图谱、软件工程.
E-mail:gaocs@ pku. edu. cn

赵 文 男,1967 年出生,辽宁大连人. 现为北京大学软件工程国家工程研究中心研究员、博士生导师,主要研究领域为知识图谱、软件工程、软件安全.
E-mail:zhaowen@ pku. edu. cn