doi:10.3969/j.issn.1672-1292.2021.04.004

基于深度学习的中文零代词识别

王立凯[1],曲维光[1,2],魏庭新[3],周俊生[1],顾彦慧[1],李　斌[2]

(1.南京师范大学计算机与电子信息学院,江苏 南京 210023)
(2.南京师范大学文学院,江苏 南京 210097)
(3.南京师范大学国际文化教育学院,江苏 南京 210097)

[摘要]　针对中文零代词识别任务,提出了一种基于深度神经网络的中文零代词识别模型. 首先,通过注意力机制利用零代词的上下文来帮助表示缺省的语义信息. 然后,利用 Tree-LSTM 挖掘零代词上下文的句法结构信息. 最后,利用语义信息和句法结构信息的融合特征识别零代词. 实验结果表明,相对于以往的零代词识别方法,该方法能够有效提升识别效果,在中文 OntoNotes5.0 数据集上的 $F1$ 值达到 63.7%.

[关键词]　深度学习,中文零指代,零代词识别,Tree-LSTM,注意力机制
[中图分类号]TP391　[文献标志码]A　[文章编号]1672-1292(2021)04-0019-08

Identification of Chinese Zero Pronouns Based on Deep Learning

Wang Likai[1],Qu Weiguang[1,2],Wei Tingxin[3],Zhou Junsheng[1],Gu Yanhui[1],Li Bin[2]

(1.School of Computer and Electronic Information,Nanjing Normal University,Nanjing 210023,China)
(2.School of Chinese Language and Literature,Nanjing Normal University,Nanjing 210097,China)
(3.International College for Chinese Studies,Nanjing Normal University,Nanjing 210097,China)

Abstract:To solve the task of Chinese zero pronoun identification, this paper proposes a Chinese zero pronoun identification model based on deep neural network. Firstly,attention mechanism is applied to learn more semantic information from the context of zero pronoun. Then,Tree-LSTM is used to capture syntactic structure features of the context of the zero pronoun. Finally,semantic information and syntactic structure information are combined to identify the zero pronoun. Compared with the previous zero pronoun identification methods,experiments on Chinese OntoNotes5.0 corpus show that our proposed approach can more effectively improve the recognition effect,and the $F1$ value reaches 63.7%.

Key words:deep learning,Chinese zero pronoun,zero pronoun identification,Tree-LSTM,attention

零指代是语言学中一种特殊的语言现象,是指为了保证语言的连贯性而省略的、且可通过上下文推断出的语言单元. 该省略的语言单元在句子中承担相应的句法成分,可以和前文中一个或多个名词短语等语言单元构成指代关系. 这种被省略的语言单元称为零代词,前文中与其构成指代关系的语言单元被称为先行语.

例句 1　虽然[卓伟]的新闻装备落后,但 Φ1 说起娱乐新闻观念,Φ2 却比许多年轻同事更自由,更激进.

例句 1 中,Φ2 就是一个零代词,指向前文中的人名“卓伟”. 但 Φ1 就不是一个零代词,前文中没有语言单元与其构成指代关系. 因此,零代词的两个必要条件是:(1)句中存在句法成分缺省;(2)前文中有语言单元与其构成指代关系. 与英文相比,汉语是一种意合型语言,其特点是形态不完整,这使得汉语中存在大量的缺省. 据 Kim[1] 统计,汉语中存在句法成分省略的现象高达 36%,而英文中的省略现象则不超过 4%,所以零指代的识别与消解对于中文信息处理来说显得更加迫切和重要.

中文零指代通常可分为两个子任务,一是零代词的识别,二是零代词的消解. 目前的研究工作多数聚焦于消解部分,并取得了丰硕成果[2-4],而零代词的识别研究相对较少. 零代词识别是指识别出句子中有回指的缺省语言单元的位置,是零指代消解必要的前期基础. 在早期的零代词识别研究中,主要是利用规

收稿日期:2021-01-19.
基金项目:国家自然科学基金项目(61772278、61472191)、国家社科基金项目(18BYY127)和江苏高校哲学社会科学优秀创新团队建设项目.
通讯作者:曲维光,博士,教授,博士生导师,研究方向:自然语言处理. E-mail:wgqu_nj@163.com

则的方法去识别零代词[5],缺点是难以制定较为全面的规则且可泛化性差. 也有一些学者如 Zhao[6]、Kong[7]等利用传统机器学习方法去识别中文零代词,但仍受限于特征模板和专家知识,忽略了语义特征对于识别零代词的重要作用. Zhao 等[6]第一次使用了基于机器学习的方法进行中文零代词的识别与消解. 实验过程中,在识别阶段使用了 13 个特征,在消解阶段使用了多达 26 个特征. Kong 等[7]提出了一个基于树核支持向量机的统一中文零指代消解框架,该框架包含 3 个部分,分别为候选零代词的生成、回指零代词的识别和消解. Chen 等[8]也提及了其使用的基于规则的零代词识别方法来作为消解工作的铺垫,并在之前研究的基础上丰富了零代词特征,利用 SVMlight 算法取得了当时最好的效果[9]. 也有学者如 Liu 等[10]提出利用 GRU 网络去学习词向量中蕴含的语义信息,但忽视了句子中不同词语所蕴含的语义信息是不同的,同时也忽视了句法结构信息在识别过程中的重要作用. Kong 等[11]在使用共指链信息进行零代词识别和消解时,将零指代识别分为两个步骤:零代词候选位置生成和零代词识别,在生成零代词候选位置时使用了句法信息,在识别时则利用词汇、句法和语义特征构造了一个分类器. Song 等[12]则使用多任务学习方法,同时使用 BERT 进行编码,对零代词识别和消解进行联合学习,但由于错误传递,效果不尽如人意,联合模型 $F1$ 值均低于 40%.

本文为解决上述问题、提升零代词识别效果,提出了一种基于深度神经网络的回指零代词识别方法,首先利用注意力机制去捕获零代词上下文的语义信息,对蕴含更多语义信息的词语分配更高的权重,同时利用 Tree-LSTM 去挖掘句法结构信息,最后通过两者的融合特征识别零代词. 相比于传统识别方法需要建立繁琐的规则体系或特征模板、泛化性差等缺点,本文通过深度神经网络从输入文本中抽取抽象特征信息来识别零代词,降低了模型对繁琐的手工特征和专家知识的依赖. 在 OntoNotes5.0 中文语料上的实验结果证明,该方法有效提升了中文零代词的识别效果.

1　模型

1.1　模型概述

本文的整体模型框架如图 1 所示,主要由 4 个部分组成:零代词候选集合的生成模块、注意力机制模块、Tree-LSTM 模块和输出模块. 首先,通过事先制定的句法规则从原始语料中筛选出零代词的候选集合. 由于零代词在文本中缺省,无法直接获取零代词的词向量,本文通过两种方式间接对其进行特征表示:一是通过注意力机制模块,包含 Encoder 层和 Attention 层,Encoder 层对零代词的上下文进行前向后向的 LSTM 编码,得到其上下文的隐藏层表示,再利用 Attention 层捕获其深层语义信息;二是通过 Tree-LSTM 模块挖掘其句法结构信息. 最后,将两者的融合特征送入输出层,得到分类结果.

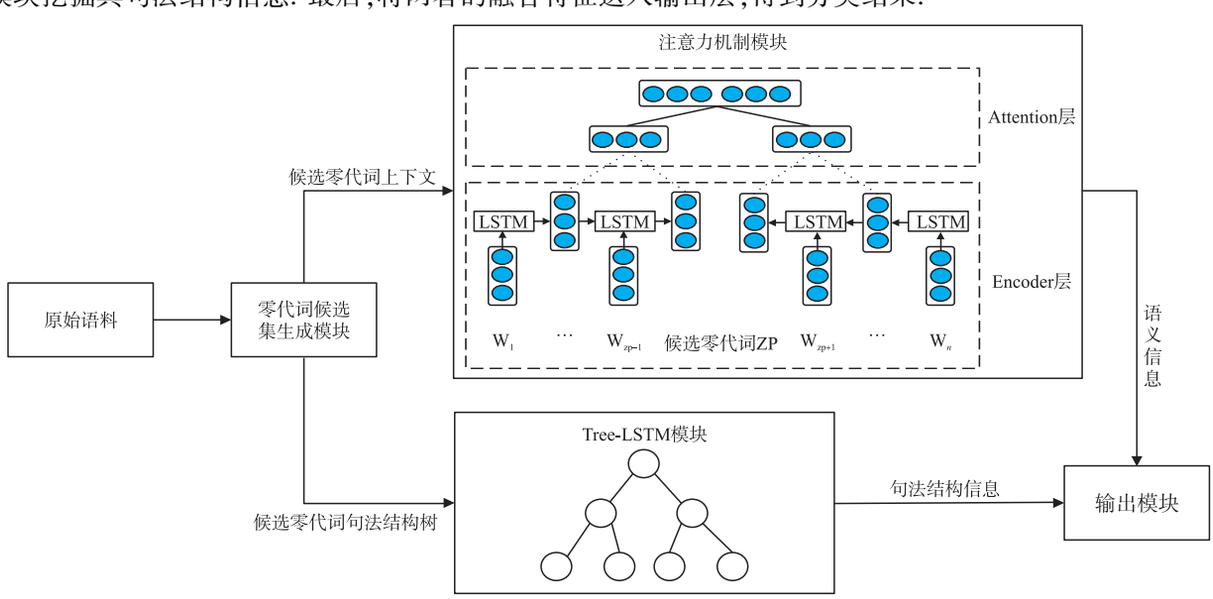

图 1　中文零代词识别整体框架图

Fig. 1　Framework of Chinese zero pronoun recognition

1.2 零代词候选集合生成模块

零指代识别和消解任务常采用 OntoNotes5.0 语料[13]数据集,本文亦采用该语料. 由于 OntoNotes5.0 语料只标记了主语位置的零代词,故本文研究只考虑主语位置零代词的情况. 现代汉语中,主语位置零代词总是出现在谓词短语节点(VP)之前,而谓词短语是由一个或一个以上动词短语构成,由此,对于句法树中的节点 T,若同时满足以下 3 条句法规则,则节点 T 的左边相邻位置是零代词的候选位置:

(1)T 节点属于 VP 节点;

(2)T 节点的父亲节点不属于 VP 节点;

(3)T 节点的左兄弟节点不属于 NP 节点,或者 T 节点没有左兄弟节点.

例句 2 建筑是开发浦东的一项主要经济活动,这些年有数百家建筑公司,四千余个建筑工地遍布在这片热土.

如图 2 所示,例句 2 中,根据上文所制定的规则,节点(VP(VV(开发)NP(NR(浦东))))的当点节点为 VP 节点,其父亲节点为 IP 节点,且无左兄弟节点,因此判定其为一个零代词候选位置,但实际上该位置并不存在零代词,因为其不满足回指前文中语言单元的条件. 而树中“这些年有数百家公司”的句首位置也满足规则,同时回指前文中的“浦东”,因此该候选位置存在一个零代词.

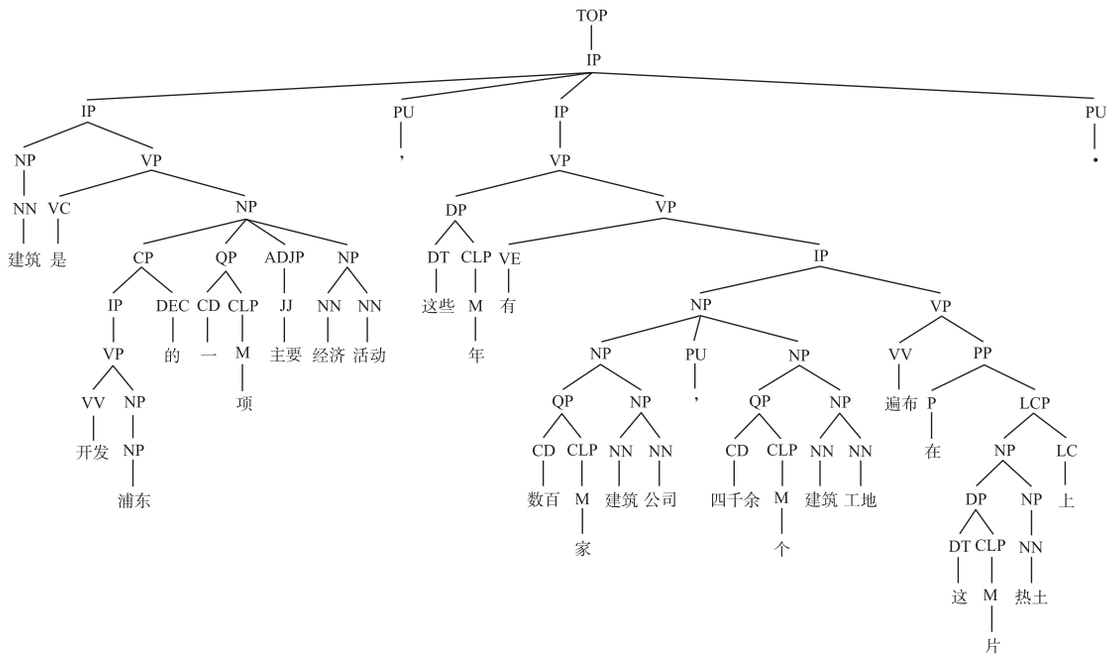

图 2 例句 2 句法分析树

Fig. 2 Syntactic parsing tree of sentence 2

1.3 注意力机制模块

1.3.1 Encoder 层

通过零代词候选集合生成模块的筛选,获取了零代词候选集合 ZPS,对于候选集合中的每一个候选零代词 zp,其上下文可表示为:

$$\boldsymbol{Context}_{\text{precedding}}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{\text{zp}-1}), \tag{1}$$

$$\boldsymbol{Context}_{\text{following}}=(\boldsymbol{x}_{\text{zp}+1},\boldsymbol{x}_{\text{zp}+2},\cdots,\boldsymbol{x}_n). \tag{2}$$

式中,$\boldsymbol{x}_i=[\boldsymbol{w}_i;\boldsymbol{p}_i]$,$\boldsymbol{w}_i$ 表示第 i 个词语的词向量,$\boldsymbol{p}_i$ 表示其词性向量,$\boldsymbol{x}_i$ 为 $\boldsymbol{w}_i$ 和 $\boldsymbol{p}_i$ 的拼接. 为了获取其上下文信息,本文采用 LSTM 网络对两个序列 $\boldsymbol{Context}_{\text{precedding}}$ 和 $\boldsymbol{Context}_{\text{following}}$ 分别进行编码. LSTM 的公式如下所示:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}^{(i)}\cdot[\boldsymbol{x}_t;\boldsymbol{h}_{t-1}]+\boldsymbol{b}^{(i)}), \tag{3}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}^{(f)}\cdot[\boldsymbol{x}_t;\boldsymbol{h}_{t-1}]+\boldsymbol{b}^{(f)}), \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}^{(o)}\cdot[\boldsymbol{x}_t;\boldsymbol{h}_{t-1}]+\boldsymbol{b}^{(o)}), \tag{5}$$

$$\widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}^{(c)}\cdot[\boldsymbol{x}_t;\boldsymbol{h}_{t-1}]+\boldsymbol{b}^{(c)}), \tag{6}$$

$$\boldsymbol{C}_t=\boldsymbol{i}_t\cdot\widetilde{\boldsymbol{C}}_t+\boldsymbol{f}_t\cdot\boldsymbol{C}_{t-1},\tag{7}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\cdot\tanh(\boldsymbol{C}_t).\tag{8}$$

式中,·代表元素乘法,$\boldsymbol{W}^{(i)}$,$\boldsymbol{b}^{(i)}$,$\boldsymbol{W}^{(f)}$,$\boldsymbol{b}^{(f)}$,$\boldsymbol{W}^{(o)}$,$\boldsymbol{b}^{(o)}$,$\boldsymbol{W}^{(c)}$和$\boldsymbol{b}^{(c)}$代表 LSTM 网络的参数.

本文将零代词的上下文通过 LSTM 网络进行编码:

$$\boldsymbol{h}_t^{\text{pre}}=\text{LSTM}_{\text{pre}}(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}^{\text{pre}}),\tag{9}$$

$$\boldsymbol{h}_t^{\text{fol}}=\text{LSTM}_{\text{fol}}(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}^{\text{fol}}).\tag{10}$$

式中,LSTM_{pre}和LSTM_{fol}分别代表一个前向的 LSTM 网络和一个后向的 LSTM 网络. 可得零代词上下文中每一个词语的隐藏层表示:

$$\boldsymbol{H}_{\text{pre}}=\{\boldsymbol{h}_1^{\text{pre}},\boldsymbol{h}_2^{\text{pre}},\cdots,\boldsymbol{h}_{n_{\text{pre}}}^{\text{pre}}\},\tag{11}$$

$$\boldsymbol{H}_{\text{fol}}=\{\boldsymbol{h}_1^{\text{fol}},\boldsymbol{h}_2^{\text{fol}},\cdots,\boldsymbol{h}_{n_{\text{fol}}}^{\text{fol}}\}.\tag{12}$$

1.3.2　Attention 层

由于零代词在文章中是缺省的,在以往的研究中,虽然学者们使用了许多方法去表示零代词,但仍存在无法获取部分关键语义信息等问题. 本文采用注意力机制去编码零代词的语义信息.

由于在 Encoder 层已经获得了零代词上下文的隐藏层表示$\boldsymbol{H}_{\text{pre}}$和$\boldsymbol{H}_{\text{fol}}$,接下来需将$\boldsymbol{H}_{\text{pre}}$和$\boldsymbol{H}_{\text{fol}}$作为注意力机制的输入,并计算权重:

$$\boldsymbol{A}_{\text{pre}}=\text{softmax}(\boldsymbol{W}_2^{\text{pre}}\cdot\tanh(\boldsymbol{W}_1^{\text{pre}}\cdot\boldsymbol{H}_{\text{pre}}^{\text{T}})),\tag{13}$$

$$\boldsymbol{A}_{\text{fol}}=\text{softmax}(\boldsymbol{W}_2^{\text{fol}}\cdot\tanh(\boldsymbol{W}_1^{\text{fol}}\cdot\boldsymbol{H}_{\text{fol}}^{\text{T}})).\tag{14}$$

式中,$\boldsymbol{W}_1$是一个 $a\times u$ 的参数矩阵,$\boldsymbol{W}_2$为 $r\times a$ 的参数矩阵,u 代表零代词上下文隐藏层输出的维度,r 代表选择的注意力机制层数.

通过该方法即可得到一个多层的权重矩阵$\boldsymbol{A}$,与单层的注意力机制权重矩阵相比,多层注意力机制使得模型可从不同的角度关注句子的不同部分,可更高效地从语义角度对零代词进行句子级信息的建模. 当得到权重矩阵$\boldsymbol{A}$后,将得到的 r 层权重与零代词的上下文隐藏层输出$\boldsymbol{H}$作加权求和:

$$\boldsymbol{M}_{\text{pre}}=\boldsymbol{A}_{\text{pre}}\cdot\boldsymbol{H}_{\text{pre}},\tag{15}$$

$$\boldsymbol{M}_{\text{fol}}=\boldsymbol{A}_{\text{fol}}\cdot\boldsymbol{H}_{\text{fol}}.\tag{16}$$

而后将每个矩阵(即$\boldsymbol{M}_{\text{pre}}$和$\boldsymbol{M}_{\text{fol}}$)的行向量取平均得到零代词的上文表示和下文表示,将两个表示向量做拼接得到零代词的语义表示向量. 通过注意力机制学习上下文中不同词对表示零代词语义信息的重要程度,完成注意力权重的分配,可更好地从语义角度对零代词进行表示.

1.4　Tree-LSTM 模块

Kong 等[7]指出,句法结构信息对于中文零代词的识别具有重要的作用. 本节通过 Tree-LSTM[14]对零代词的成分句法树进行编码,从结构化句法信息的角度对零代词进行更好地表示. 传统的 LSTM 模型在每个时间步都是将当前时间步的词语信息和上一个时间步的输出信息作为输入,更新当前时间步的记忆单元和隐藏层状态. 在 Tree-LSTM 中,树中每个节点的输入为当前节点的词语信息和其孩子节点的输出信息,而不只考虑前一步的信息. 给定一个成分句法树,对于节点 j,Tree-LSTM 的具体公式如下:

$$\boldsymbol{i}_j=\sigma\left(\boldsymbol{W}^{(i)}\boldsymbol{x}_j+\sum_{l=1}^{N}\boldsymbol{U}_l^{(i)}\boldsymbol{h}_{jl}+\boldsymbol{b}^{(i)}\right),\tag{17}$$

$$\boldsymbol{f}_{jk}=\sigma\left(\boldsymbol{W}^{(f)}\boldsymbol{x}_j+\sum_{l=1}^{N}\boldsymbol{U}_{kl}^{(f)}\boldsymbol{h}_{jl}+\boldsymbol{b}^{(f)}\right)_{k=1,\cdots,N},\tag{18}$$

$$\boldsymbol{o}_j=\sigma\left(\boldsymbol{W}^{(o)}\boldsymbol{x}_j+\sum_{l=1}^{N}\boldsymbol{U}_l^{(o)}\boldsymbol{h}_{jl}+\boldsymbol{b}^{(o)}\right),\tag{19}$$

$$\boldsymbol{u}_j=\tanh\left(\boldsymbol{W}^{(u)}\boldsymbol{x}_j+\sum_{l=1}^{N}\boldsymbol{U}_l^{(u)}\boldsymbol{h}_{jl}+\boldsymbol{b}^{(u)}\right),\tag{20}$$

$$\boldsymbol{C}_j=\boldsymbol{i}_j\cdot\boldsymbol{u}_j+\sum_{l=1}^{N}\boldsymbol{f}_{jl}\cdot\boldsymbol{c}_{jl},\tag{21}$$

$$\boldsymbol{h}_j=\boldsymbol{o}_j\cdot\tanh(\boldsymbol{c}_j).\tag{22}$$

式中,$\boldsymbol{x}_j=[\boldsymbol{w}_j;\boldsymbol{p}_j;\boldsymbol{t}_j]$,$\boldsymbol{w}_j$为节点 j 的词向量,$\boldsymbol{p}_j$为节点 j 的词性向量,$\boldsymbol{t}_j$为节点 j 的句法标签向量,$\boldsymbol{x}_j$为三者

的拼接;$\boldsymbol{h}_{jl}$表示节点 j 的第 l 个子节点的隐藏层状态,N 表示节点 j 的孩子节点数(在应用到成分句法树时,本文将句法树转化为二叉句法树[15],所以 N 为 2);σ 表示 sigmoid 函数;$\boldsymbol{f}_{jk}$表示节点 j 第 k 个遗忘门;$\boldsymbol{C}_j$ 为当前节点 j 的记忆单元,$\boldsymbol{c}_{jl}$为结点 j 第 l 个孩子的记忆单元;$\boldsymbol{h}_j$ 为当前节点 j 最终的输出. 通过这种做法,模型能够学习到对子节点状态更细粒度的条件集成,例如对于句法树中的一个节点,其左孩子对应名词短语,右孩子对应动词短语,假设需要更加强调动词短语,则式(18)中参数 $\boldsymbol{U}_{kl}^{(f)}$ 的训练结果会使 f 左接近 0(倾向于遗忘左子节点信息)而 f 右接近 1(倾向于保留右子节点信息). 此外,为了减少句法树中无关节点对零代词识别效果的影响,本文对零代词的句法结构树进行了裁剪:

(1)保留句法树根节点到零代词出现位置动词短语节点的路径,并保留此路径上直接相连的名词短语节点和动词短语节点;

(2)保留零代词出现位置前一个动词短语节点和零代词出现位置动词短语之间的句法树;

(3)对于动词短语节点的子树,只保留那些以动词和名词为叶节点的路径.

对图 2 句法树中"开发浦东"前的候选零代词进行裁剪后的句法树结构如图 3 所示.

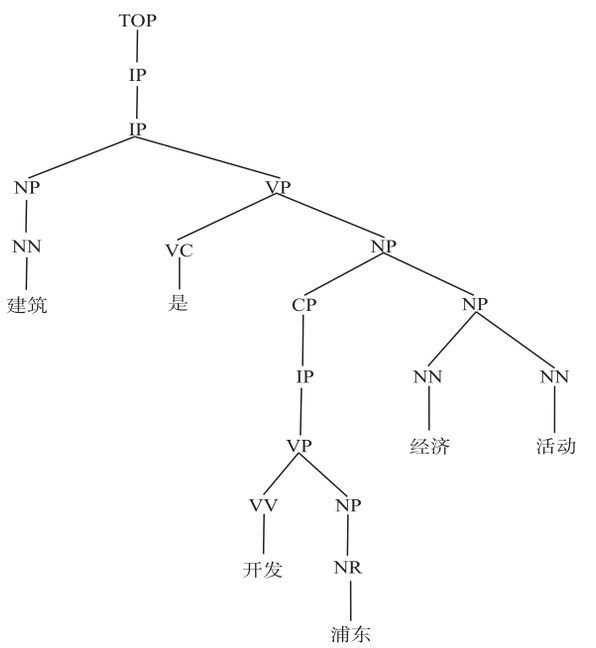

图 3　候选零代词的裁剪句法树

Fig. 3　Pruned syntactic tree of candidate zero pronoun

1.5　输出模块

通过注意力机制模块和 Tree-LSTM 模块分别获取候选零代词语义层次和句法层次的表示后,本文将其送入一个两层前馈神经网络中,判断其是否为一个零代词. 计算公式如下:

$$\boldsymbol{v}_a=[\mathrm{avg}(\boldsymbol{M}_{\mathrm{pre}});\mathrm{avg}(\boldsymbol{M}_{\mathrm{fol}})], \tag{23}$$

$$\boldsymbol{s}_1=\tanh(\boldsymbol{W}_1\cdot[\boldsymbol{v}_a;\boldsymbol{h}_{\mathrm{pre}};\boldsymbol{h}_{\mathrm{fol}};\boldsymbol{v}_t;\boldsymbol{v}_{\mathrm{fea}}]+\boldsymbol{b}_1), \tag{24}$$

$$\boldsymbol{s}_2=\tanh(\boldsymbol{W}_2\cdot\boldsymbol{s}_1+\boldsymbol{b}_2). \tag{25}$$

式中,$\boldsymbol{v}_a$ 表示注意力机制模块中 Attention 层所得矩阵 $\boldsymbol{M}_{\mathrm{pre}}$ 和 $\boldsymbol{M}_{\mathrm{fol}}$ 行向量取平均的拼接向量;$\boldsymbol{h}_{\mathrm{pre}}$ 和 $\boldsymbol{h}_{\mathrm{fol}}$ 分别表示注意力机制模块中 Encoder 层中 LSTM 前向和后向的最终隐层状态;$\boldsymbol{v}_t$ 表示零代词句法裁剪树经过 Tree-LSTM 模块编码后的输出向量;$\boldsymbol{v}_{\mathrm{fea}}$为候选零代词句法结构、位置等手工构建的特征信息所组成的数值向量,具体特征描述与文献[3]中一致;$\boldsymbol{W}_1,\boldsymbol{b}_1,\boldsymbol{W}_2,\boldsymbol{b}_2$ 分别代表两层前馈神经网络的参数;$\boldsymbol{s}_1$ 和 $\boldsymbol{s}_2$ 分别代表两层前馈神经网络的输出.

最后,将其输出送入 softmax 层得到最终的分类概率为:

$$\boldsymbol{P}=\mathrm{softmax}(\boldsymbol{W}_{\mathrm{soft}}\cdot\boldsymbol{s}_2+\boldsymbol{b}_{\mathrm{soft}}), \tag{26}$$

式中,$\boldsymbol{P}$ 表示候选零代词是回指零代词的概率;$\boldsymbol{s}_2$ 为前两层隐藏层的输出.

1.6　损失函数

本文采用交叉熵作为模型训练的目标函数. 损失函数的公式为:

$$\mathrm{loss}=-\sum_{z\in Z}\sum_{i\in C}y_i(z)\log p_i(z), \tag{27}$$

式中,Z 为零代词候选集;$p_i(z)$表示通过本文模型的候选零代词 z 属于类别 i 的概率;$y_i(z)$为候选零代词 z 的标签.

2　实验

2.1　评测指标

与早期关于中文零指代任务的研究相似,本文通过召回率(Recall)、精确率(Precision)和 F 值(F-score)来评估本文的模型,其中 Recall 和 Precision 定义为:

$$R_{\mathrm{Recall}}=\frac{\mathrm{AZP}_{\mathrm{right}}}{\mathrm{AZP}_{\mathrm{gold}}}, \tag{28}$$

$$P_{\text{Precision}}=\frac{\text{AZP}_{\text{right}}}{\text{AZP}_{\text{res}}}, \tag{29}$$

$$F=\frac{2*P_{\text{Precision}}*R_{\text{Recall}}}{P_{\text{Precision}}+R_{\text{Recall}}}. \tag{30}$$

式中,$\text{AZP}_{\text{right}}$表示本文模型预测正确的零代词;$\text{AZP}_{\text{res}}$表示本文模型预测为正确的所有零代词;$\text{AZP}_{\text{gold}}$表示语料中标注的所有零代词.

2.2　数据集

实验数据选择 CoNLL-2012 共享任务中的 OntoNotes5.0 数据集[13],仅选用其中中文部分. 中文部分数据集只有在训练集和验证集上才有零代词的标注,在测试集上未作标记. 本文参照以往中文零指代文献中的做法,将模型在训练集上进行训练,在验证集上进行评测. OntoNotes5.0 中文数据集主要包含 6 种类别,分别为广播会话、广播新闻、杂志、通讯新闻、电话对话和博客文章. 数据集的具体统计数据如表 1 所示.

表 1　OntoNotes5.0 数据集统计

Table 1　Statistics of the OntoNotes 5.0 dataset

类别	训练集	测试集
文档数	1 391	172
句子数	36 487	6 083
单词数	756 063	110 034
零代词	12 111	1 711

2.3　参数设置

由于实验缺少验证集,本文从训练语料中切分了约 20%的数据作为验证集,用以调整超参数. 具体参数设置如表 2 所示.

表 2　实验超参数设置

Table 2　Parameter setting of the experiments

超参数名称	参数值	超参数名称	参数值
词语向量维数	100	Dropout	0.5
词性向量维数	50	学习率	0.003
LSTM 隐藏层维数	256	注意力机制层 a 维数	128
句法标签向量维数	50	前馈网络隐藏层 1 维数	256
Tree-LSTM 隐藏层维数	256	前馈网络隐藏层 2 维数	512

在实验过程中发现,由于训练集正负样本不平衡(正负比例约为 1∶3),导致实验效果下降. 因此,本文通过复制正例样本来平衡训练集中正负样本的比例. 经测试,最终选择将正例样本复制 3 次,模型达到了最优的效果.

本文将模型初始迭代次数设置为 50. 图 4 所示为本文模型在验证集上的学习曲线,可以观察到模型在第一个 epoch 时,验证数据集上的 F1-score 约为 56%,随着迭代次数的增加,F1-score 值不断增加. 在第 11 次迭代时,模型在验证集上的性能达到最优. 之后,本文模型的 F1-score 开始下降. 因此本文保存验证集上效果最优的模型参数,在测试集上进行评价.

此外,本文希望多层次的句子级信息可以为表示零代词提供丰富的语义信息,因而通过改变注意力机制层的 r 值,来评估不同的 r 值对于中文零代词识别的影响. 当 $r=16$ 的时候,模型达到最优值,如图 5 所示. 本文通过注意力机制将模型关注的重点放在对编码零代词更加有用的部分,将句子以零代词为中心分成两个部分,当 $r=16$ 时,说明本文的模型分别从 16 个不同的角度去编码前半部分和后半部分的语义信息,因而能够达到较好的结果.

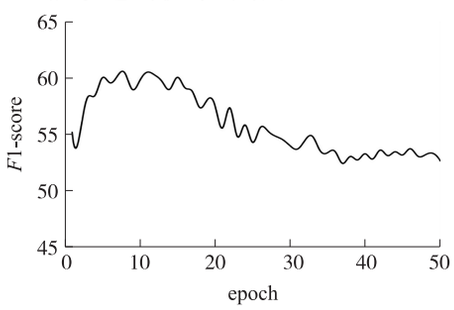

图 4　模型在验证数据集上的学习曲线

Fig. 4　Learning curve of the model on the validation dataset

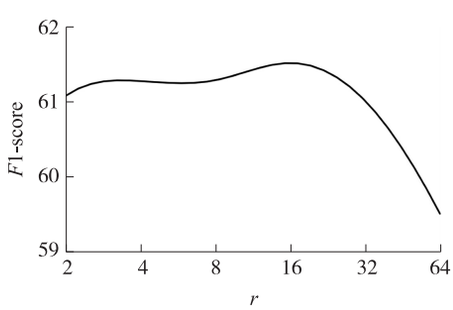

图 5　r 值对零代词识别的影响

Fig. 5　Effect of r value on zero pronoun recognition

2.4 实验结果与分析

表3所示为本文模型和近年中文零代词识别任务在OntoNotes5.0中文语料上的实验结果的对比,前4行为以往对于中文零代词识别任务的实验结果,后两行为本文的实验结果. 可以看出,本文模型比以往最优模型的F值要高出1.4%,加入Bert预训练词向量后的F值更是高出了3.6%.

表3 不同模型实验结果对比

Table 3 Comparison of experimental results of different models

方法	P/%	R/%	F/%	方法	P/%	R/%	F/%
Chen and Ng(2013)	55.1	50.6	52.8	Liu et al. (2017)	42.8	73.3	54.1
Chen and Ng(2014)	42.3	72.4	53.4	Rules+Attention+Tree-LSTM	49.9	80.2	61.5
Chen and Ng(2016)	50.1	75.1	60.1	Rules+Bert+Attention+Tree+LSTM	54.6	76.4	63.7

从精确率来看,本文模型在未加入Bert之前比表中Chen and Ng(2013)和(2016)的模型分别低了5.2%和0.2%. 分析发现,这两个模型均采用传统机器学习的方法. 中文零代词具有较强的规则性,制定合适的特征模板会对中文零代词识别的精确率具有较好的正向作用. 从召回率来看,本文的模型无论是加Bert之前还是之后,均优于以往的中文零代词识别模型.

由于近期Bert模型在各大任务上均取得了较好的成果,本文也尝试了在现有模型上加入Bert预训练模型,主要是利用Bert的encoder层输出去增强本文模型中的词向量表示. 由于Bert的pytorch版本目前尚只能做字符级编码,本文对中文词语中每个字符的向量表示做了均值操作,以此来代替词级向量. 本文并未对Bert网络进行微调,仅将Bert网络的输出作为特征. 实验表明,加入Bert后模型在OntoNote5.0中文语料上的F值、准确率和召回率分别为63.7%、54.6%和76.4%,F值比原始模型高出了2.2%,精确率上升了4.7%,但召回率下降了3.8%.

2.5 模块分析

表4所示为本文中各个模块在模型中所起的作用.

表4 各模块在实验中的作用对比

Table 4 Comparison of experimental results with different modules

方法	P/%	R/%	F/%	方法	P/%	R/%	F/%
Rules+Attention+Tree-LSTM	49.9	80.2	61.5	Rules+Attention	47.2	73.6	57.5
Rules-based	24.2	99.1	38.9	Rules+Zpcenter	45.4	76.0	56.8
Rules+Tree-LSTM	46.7	84.7	60.2				

如表4所示,Rules-based表示本文零代词候选集生成模块所生成零代词候选集的实验效果. 在本文的规则下,回指零代词的召回率很高,达到99.1%,但是精确率较低,只有24.2%. 因为本文主要是利用规则去生成零代词的候选集合,对召回率的要求更高.

Rules+Tree-LSTM表示本文模型去除注意力机制模块之后的实验结果,其F值为60.2%,精确率为46.7%,召回率为84.7%. 由于成分句法树的成分缺省具有较强的规律性,通过Tree-LSTM对成分句法树的学习,理论上可以找回绝大多数零代词的缺省位置. 但因语义信息的不足,对判断其是否回指存在不足.

Rules+Attention表示本文模型去除Tree-LSTM之后的实验结果,其F值为57.5%,精确率为47.2%,召回率为73.6%. 相比于Rules+Tree-LSTM,准确率略高,而召回率大幅下降,这主要是因为句法结构信息的缺失导致无法准确识别出句子中成分缺省.

Rules+Zpcenter表示同时去除Tree-LSTM模块和注意力机制模块之后的实验结果,其F值为56.8%,准确率为45.4%,召回率为76.0%. F值的再次降低证明了本文注意力机制和Tree-LSTM的有效性.

2.6 错误分析

对模型输出的错误结果进行分析(例如句3)分析发现,零代词回指的名词短语虽大部分都在同一句话内,但也有部分零代词所回指的名词短语是在之前句子中出现的,表5给出了相关的统计数据. 本文的模型主要是在单句内进行,与前文中的句子并无联系.

表5 与先行词不在同一句内的零代词数量统计

Table 5 Statistics of zero pronouns that are not in same sentence with antecedants

类别	训练集	测试集
零代词总数	12 111	1 711
分离的零代词总数	4 370	636

例句 3 本届锦标赛到今天为止,已决出 7 枚金牌. * pro * 明天将进行最后 5 个男女单项的决赛.

在例句 3 中,零代词"pro"回指前一句话中的名词短语"本届竞标赛". 本文模型只能在同一句内对零代词信息进行捕捉,对跨句信息暂无法获取.

3 结论

本文针对中文零代词识别任务,利用 LSTM 编码中文零代词的上下文信息,再通过注意力机制捕捉句子中的关键语义,有效利用上下文语义,减少上下文无关信息对零代词表示的影响. 使用 Tree-LSTM 对零代词上下文的句法结构信息进行编码,将语义信息和句法结构信息的融合特征通过两层前馈神经网络对零代词进行识别. 本文首次通过深度学习方法同时利用语义信息和句法结构信息进行中文零代词的识别,实验结果表明,该方法在 OntoNotes5.0 语料上取得了较好的效果.

[参考文献](References)

[1] KIM Y J. Subject/object drop in the acquistion of korean:a cross-linguistic comparison[J]. Journal of East Asian Linguistics, 2000,9(4):325-351.

[2] YIN Q Y,ZHANG Y,ZHANG W N,et al. Deep reinforcement learning for Chinese zero pronoun resolution[C]//Proceedings of 56th Annual Meeting of the Association for Computational linguistics. Melbourne:Association for Computational Linguistics, 2018:569-578.

[3] YIN Q Y,ZHANG Y,ZHANG W N,et al. Zero pronoun resolution with attention-based neural network[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe:Association for Computational Linguistics,2018:13-23.

[4] LIN P Q,YANG M. Hierarchical attention network with pairwise loss for Chinese zero pronoun resolution[C]//Proceedings of 34th AAAI Conference on Artificial Intelligence. Palo Alto:AAAI Press,2020:8352-8358.

[5] 秦凯伟,孔芳,李培峰,等. 基于规则的中文零指代项识别研究[J]. 计算机科学,2012,39(10):278-281.

[6] ZHAO S H,HWEE T N. Identification and resolution of Chinese zero pronouns:a machine learning approach[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague:Association for Computational Linguistics,2007:541-550.

[7] KONG F,ZHOU G D. A tree kernel-based unified framework for Chinese zero anaphora resolution[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge:Association for Computational Linguistics,2010: 882-891.

[8] CHEN C,NG V. Chinese zero pronoun resolution:an unsupervised approach combining ranking and integer linear programming [C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec:AAAI Press,2014:1622-1628.

[9] CHEN C,NG V. Chinese zero pronoun resolution with deep neural networks[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin:Association for Computational Linguistics,2016:778-788.

[10] LIU B Q,DU X K,LIU M,et al. Resolving Chinese zero pronoun with word embedding[C]//National CCF Conference on Natural Language Processing and Chinese Computing. Dalian:Springer,2017:828-838.

[11] KONG F,ZHOU G D. Chinese zero pronoun resolution:a chain to chain approach[C]//National CCF Conference on Natural Language Processing and Chinese Computing. Dalian:Springer,2017:393-405.

[12] SONG L F,XU K,ZHANG Y,et al. ZPR2:joint zero pronoun recovery and resolution using multi-task learning and BERT[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online:Association for Computational Linguistics,2020:5429-5434.

[13] PRADHAN S,MOSCHITTI A,XUE N W,et al. CoNLL-2012 shared task:modeling multilingual unrestricted coreference in OntoNotes[C]//Proceedings of the Joint Conference on EMNLP and CoNLL:Shared Task. Stroudsburg:Association for Computational Linguistics,2012:1-40.

[14] TAI K S,SOCHER R,MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing:Association for Computational Linguistics,2015:1556-1566.

[15] ERIGUCHI A,HASHIMOTO K,TSURUOKA Y. Tree-to-sequence attentional neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin:Association for Computational Linguistics, 2016:823-833.

[责任编辑:严海琳]