



结合实体共现信息与句子语义特征的关系抽取方法

马语丹, 赵义, 金婧, 万怀宇*

北京交通大学计算机与信息技术学院交通数据分析与挖掘北京市重点实验室, 北京 100044

* 通信作者. E-mail: hywan@bjtu.edu.cn

收稿日期: 2018-10-05; 接受日期: 2018-10-30; 网络出版日期: 2018-11-14

国家重点研发计划 (批准号: 2018YFC0830200) 资助项目

摘要 实体关系抽取是信息抽取领域的重要任务之一, 也是知识图谱构建的一个关键环节. 现有的关系抽取方法大多都是围绕实体对从句子中抽取上下文语义特征, 然后进行关系分类, 这忽略了实体在整个语料集中的全局上下文特征. 本文提出了一种新颖的结合实体共现信息与句子语义信息的神经网络 (CNSSNN) 模型, 用于实体关系抽取. 该模型首先构造整个语料集蕴含的实体共现关系网络, 并通过引入注意力机制有侧重地提取实体的网络环境信息, 从而为各个实体生成语料级全局上下文特征, 同时利用双向门控循环单元网络 (bi-GRU) 为实体对提取句子级上下文语义特征, 最后将语料级特征和句子级特征结合起来, 进行实体关系抽取. 在公开数据集和人工标注的数据集上的实验结果表明, 本文提出的方法其准确率和召回率要明显优于其他现有方法.

关键词 信息抽取, 实体关系抽取, 实体共现网络, 注意力机制, 门控循环单元

1 引言

实体关系抽取作为信息抽取领域的一个十分重要的研究问题, 其任务是通过语料中的上下文信息, 抽取出具有相关关系的实体对 (entity pair), 并进一步确认该实体对的关系类型^[1]. 关系抽取的结果可以直接用于知识图谱中实体关系网络的构建, 也可以应用于机器翻译、问答系统等复杂的自然语言处理任务^[2]. 关系抽取结果的好坏, 将直接对其后续任务的效果产生重要的影响.

传统的关系抽取方法主要是从语料信息中提取出词性、句法结构、语义依存关系等表面特征和结构化特征, 并用模式匹配^[3,4]、支持向量机 (SVM)^[5,6]、条件随机场 (CRF)^[7] 等算法对实体对之间的关系进行分类. 如何设计更有效的特征是传统关系抽取方法的核心, 一个好的特征设计往往能取得很好的抽取结果. 传统关系抽取方法中的特征设计和组合往往对语料有很强的针对性, 特征提取的结果

引用格式: 马语丹, 赵义, 金婧, 等. 结合实体共现信息与句子语义特征的关系抽取方法. 中国科学: 信息科学, 2018, 48: 1533–1545, doi: 10.1360/N112018-00157
Ma Y D, Zhao Y, Jin J, et al. Combining entity co-occurrence information and sentence semantic features for relation extraction (in Chinese). Sci Sin Inform, 2018, 48: 1533–1545, doi: 10.1360/N112018-00157

也会受到自然语言处理 (NLP) 工具的影响, 并在关系抽取过程中产生误差传播, 因此存在针对性强、扩展性弱的缺点.

基于神经网络的关系抽取方法, 通过神经网络模型 (如 CNN [2,8~10], RNN [11~13] 等) 直接提取语料的深层语义特征用于关系分类, 不再过多依赖于人工特征的设计, 极大地减少了数据预处理过程中对 NLP 工具的依赖, 改善了传统关系抽取方法中误差传播、扩展性差的问题, 在关系抽取领域得到了广泛的应用. 但是, 现有的基于神经网络的关系抽取方法, 通常只从单个句子内提取语义特征, 而语料中的句子可能存在长度过短、上下文词所提供的信息较少等问题, 因此句子级 (sentence-level) 语义特征为关系抽取所提供的信息可能是有限的.

事实上, 从整个语料集的角度, 它可能包含了更丰富的对关系抽取有用的全局性信息. 具体来说, 一个实体可能不仅仅只在某一个句子中出现, 只与某一个实体存在潜在的关系. 在语料足够丰富的情况下, 一个实体可能存在于多个句子中, 与多个实体之间都存在潜在的某种关系. 那么, 将整个语料集中的全部实体的所有潜在关系提取出来, 就形成了一个完整的实体共现网络 (entity co-occurrence network). 在这个网络中, 一个实体的相邻实体集合, 就形成了它的全局性语料级 (corpus-level) 上下文信息. 例如, 实体 A 在共现网络中分别与实体 B, C, D 相邻, 则实体 B, C, D 就构成了实体 A 的语料级上下文, 可以在一定程度上从某种语义层面代表实体 A.

基于上述思想, 为了能够充分利用语料级的上下文特征来辅助关系抽取, 本文提出了一种新颖的结合实体共现网络信息与句子级语义信息的神经网络 (co-occurrence network and sentence semantics combination neural networks, CNSSNN) 关系抽取模型. 这里实体共现关系定义为任意两个实体在同一句子中的共同出现关系. 为了得到全局上下文特征, 首先从整个语料集的全部句子中提取出所有的实体共现关系, 构建一个完整的实体共现网络. 然后, 通过引入网络注意力机制 (network attention mechanism) [14], 为每个实体计算其与邻居实体之间的关系强弱, 并以此为权重, 对邻居实体的向量表示进行加权求和, 作为该实体的语料级全局特征, 该特征包含了实体对的上下文实体信息. 与此同时, 我们利用双向门控循环单元网络 (bi-GRU) 提取实体对的句子级语义特征, 该特征包含了实体对的上下文词的语义信息. 最后, 将语料级上下文特征和句子级上下文特征结合起来, 进行实体关系抽取. CNSSNN 模型可以同时捕获微观层面的句子级语义特征和宏观层面的语料级全局上下文特征, 从而弥补单个句子可能存在的句子长度太短、上下文信息太少的缺陷, 为实体关系抽取提供更丰富、更全面的分类特征.

为了评价本文提出的 CNSSNN 模型的关系抽取性能, 本文使用两个数据集进行了实验验证. 一个是来自于互联网新闻语料的人工标注数据集. 我们采用百度 NLP 分词工具¹⁾对新闻语料进行分词和实体识别后, 共标注了 9040 个“人-机构”关系, 关系标签包括“持有” (hold)、“就读” (study at)、“就业” (work at) 和“其他” (others), 共 4 类标签. 另一个数据集是语义测评数据集 SemEval-2010 中的关系分类任务数据 [15], 是一个广泛使用的标准数据集. 实验结果表明, 本文提出的 CNSSNN 模型在两个数据集上的表现都优于其他现有的关系抽取方法.

本文的主要贡献概括如下:

(1) 提出了一种结合实体共现网络信息与句子级语义信息的神经网络关系抽取模型 CNSSNN, 该模型可以分别从实体层面和词的层面, 同时提取实体对的语料级全局上下文特征和句子级局部上下文语义特征来提升关系抽取效果.

(2) 采集并标注了一个包含 9040 条“人-机构”关系的实验数据集, 可用于关系抽取任务的模型训练和测试.

1) <https://cloud.baidu.com/product/nlp>.

(3) 通过与其他现有关系抽取方法的对比实验,验证了本文所提出的方法的有效性.

本文剩余部分组织如下:第2节介绍关系抽取的相关研究;第3节详细描述本文提出的模型;第4节进行实验对比和结果分析;第5节对全文进行总结.

2 相关工作

实体关系抽取作为信息抽取的一项重要任务,一直是自然语言处理领域学者们的研究重点.早期的实体关系抽取方法多为基于特征向量的分类方法,其核心在于如何从文本语料中抽取出词法、句法、语法结构等表面特征以及结构化特征,并使用 SVM^[16]、最大熵^[17]、CRF^[18]等算法,对实体关系进行分类.例如,Zhang等^[19]将中文实体名和关系识别看做一系列分类问题,将分类过程分为学习阶段和抽取阶段,学习阶段用标注数据训练若干分类器,抽取阶段用学到的分类器抽取中文实体名和它们的关系.Suchanek等^[20]基于链接语法分析待抽取关系的模式,将模式作为分类模型 kNN 和 SVM 的特征输入,该方法验证了深层语法结构在关系识别中的重要性.Mikolov等^[21]综合考虑了实体单词、实体类型、实体引用方式、重叠、依存树和解析树等特征,实现了最大熵模型的关系分类器,该研究表明多个层次的语言学特征能够提升关系抽取的效果.Kambhatla等^[22]系统地研究了如何把包括基本词组块在内的各种特征广泛组合起来,探讨了各种语言特征对关系抽取性能的贡献,并深入研究了 WordNet 和 NameList 等语义信息对关系抽取的影响.他们的实验结果表明,基本词组块能有效提升关系抽取性能.以上这些基于特征向量的关系抽取方法往往依赖于句法分析、依存分析等 NLP 工具对数据进行预处理操作,抽取结果也较多地受到 NLP 工具处理结果的影响.

与传统的实体关系抽取方法相比,基于神经网络的关系抽取方法减少了对前期的 NLP 预处理工作的依赖,极大地弱化了 NLP 工具对抽取结果带来的影响.Zeng等^[8]利用卷积神经网络(CNN)来提取语法和句子级别的特征,并将词与中心实体在句子中的相对位置作为特征引入 CNN 的输入,对关系分类结果产生了极大促进作用.Socher等^[11]使用循环神经网络(RNN)学习两个目标名词之间的句法树路径的向量表示,并使用该向量表示进行实体关系分类.Hashimoto等^[12]在关系分类任务上,也使用了 RNN,他们提出不同的短语对不同的目标任务有明确的权重区分的想法,并通过实验验证了该想法.Santos等^[2]基于 CNN 提出了一种 Ranking-CNN 模型,并定义了一种新的损失函数,减少关系分类过程中对人工的依赖,发现在分类任务中,忽略“其他”类能大大提高准确率和召回率.Wang等^[9]在关系分类任务中提出了基于多层注意力机制的 CNN 模型,能够同时捕获实体与关系特征,从而不需要加入额外的从属结构等信息,并可以忽略句子的异构性,也取得了非常好的效果.Miwa等^[13]通过实验发现,基于长短期记忆的 RNN 能更好地提取语言结构特征,在关系分类中的表现比基于 CNN 的模型效果更好.Zhou等^[23]在 bi-LSTM 模型的基础上,通过引入注意力机制^[14]克服词的位置影响,提取句子中对关系分类任务帮助最大的重要语义信息.Ji等^[10]通过引入维基百科中的实体描述信息,来补充实体的背景特征,并通过句子层的注意力机制选择有效的训练实例,在远程监督关系抽取中取得了较好的结果.

上述现有的关系抽取方法基本上都是从单个句子里面的上下文信息中为实体关系对提取语义特征^[2,8,9,11,13],或者是加入某种外部描述信息作为实体的补充特征^[10],这些方法都忽略了语料中可能存在的句子长度较短、上下文信息不充分的问题.为了克服这一缺陷问题,本文提出一种结合实体共现网络与句子级语义信息的神经网络实体关系抽取方法,同时从宏观的实体共现角度和微观的上下文词角度,分别为实体计算语料级的全局上下文特征和句子级的局部语义特征.在语料级全局特征计算中,通过引入网络注意力机制,对各实体的邻居实体信息进行有侧重地提取.语料级全局特征和句子

表 1 相关符号及其含义
Table 1 Symbols and their description

Symbol	Description
C	The corpus
s	A sentence in corpus C
S	The matrix representation of a sentence s
w	A word in a sentence
w	The vector representation of a word w
e	An entity
f^c	Corpus-level features
f^s	Sentence-level features
f	Features of entity pair after features combination

级微观特征这两类特征相互补充, 将为实体关系抽取任务提供更丰富的特征信息.

3 结合实体共现网络与句子语义信息的神经网络模型

3.1 相关定义

设 $C = \{s_1, s_2, \dots, s_m\}$ 表示一个包含 m 个句子的语料集, 对于 $\forall s \in C$, $s = \{w_1, w_2, \dots, w_n\}$, 其中 w_i 表示句子中第 i 个词. 分别用 e_h 和 e_t 表示任意句子 s 中出现的两个实体, 则关系抽取的目标就是根据相关语义信息, 为语料集中每个句子 s 中的实体对 $\langle e_h, e_t \rangle$ 预测其关系类型.

本文中用到的相关符号及其含义如表 1 所示.

3.2 模型描述

为了同时捕获实体的语料级全局上下文特征和句子级局部上下文语义信息, 本文提出一种结合实体共现网络特征与句子语义信息的神经网络 (CNSSNN) 模型. 模型的整体框架如图 1 所示.

CNSSNN 模型一共由 3 个组件构成 (如图 1 中的虚线框所示).

(1) 语料级特征抽取组件. 以整个语料 C 为输入, 从语料中的所有句子中提取实体对之间的共现关系 (有向的). 所有的实体共现关系构成了一个整体的实体共现网络. 接着, 通过引入注意力机制为每个实体 e 的每个邻居实体计算一个关系权重, 以表征不同的邻居对实体 e 在关系抽取任务中的重要程度. 最后, 对实体 e 的所有邻居实体的词向量进行加权求和, 得到 e 的语料级上下文特征.

(2) 句子级特征抽取组件. 以句子 s 为输入, 通过双向门控循环单元 (bi-directional gated recurrent unit, bi-GRU) 网络和池化 (pooling) 操作抽取实体对 $\langle e_h, e_t \rangle$ 的句子级上下文语义特征.

(3) 分类器组件. 通过线性变换和连接操作, 融合实体 e_h 和 e_t 的语料级特征和实体对 $\langle e_h, e_t \rangle$ 的句子级语义特征, 然后采用 softmax 分类器, 为实体对 $\langle e_h, e_t \rangle$ 计算出各关系标签的概率分布.

接下来分别对模型中的词向量表示、语料级特征抽取、句子级特征抽取, 以及实体关系分类逐一进行详细的描述.

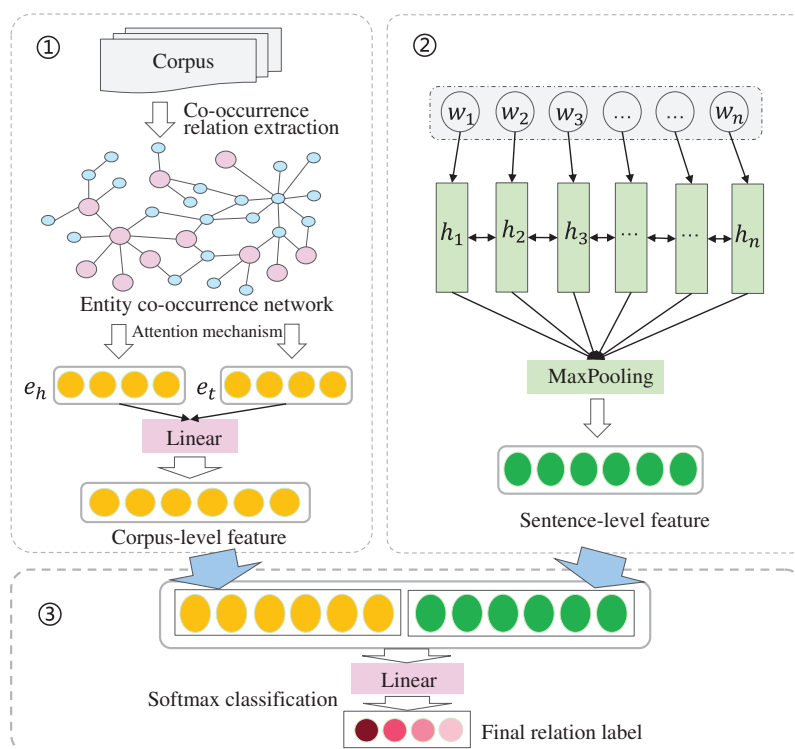


图 1 (网络版彩图) CNSSNN 模型总体框架

Figure 1 (Color online) Overall framework of CNSSNN

3.2.1 词向量表示

词向量表示用来描述一个词在语料中所代表的语义信息, 是 CNSSNN 模型的基础输入数据. 本文通过词嵌入方法^[24]对整个语料集进行训练, 得到语料的词向量矩阵 $\mathbf{L} \in \mathbb{R}^{d \times |V|}$, 其中 d 为词向量的维度, $|V|$ 为语料词典大小. 词向量矩阵中的每一列表示语料中的一个词, 每个词被表征为长度为 d 的连续实值向量 $\mathbf{w}^d$. 由此, 模型可以通过每个句子中包含的词从词向量矩阵 $\mathbf{L}$ 中得到句子的向量表示.

在关系抽取任务中, Zeng 等^[8]在 CNN 模型的基础上, 加入了词与实体在句子中的相对位置信息, 通过对句子中实体对的相对位置以及不同的词与实体之间的相对位置进行嵌入式表示, 提升了关系抽取的效果. 本文也引入词的位置嵌入信息, 将句子中每个词与实体 e 的相对位置 p 映射成一个 d_p 维的距离向量 $\mathbf{w}^p$, 用以表征句子中每个词 w 与实体 e 的相对距离, 其中 d_p 为超参.

给定句子 s 中的第 i 个词 w_i , 其词向量表示为 $\mathbf{w}_i^d \in \mathbb{R}^d$, 相对于两个实体 e_h 和 e_t 的位置向量表示为 $\mathbf{w}_i^p = [\mathbf{w}_i^{ph}; \mathbf{w}_i^{pt}] \in \mathbb{R}^{1 \times 2d_p}$, 其中 $\mathbf{w}_i^{ph}$ 是 w_i 与实体 e_h 之间的距离向量, $\mathbf{w}_i^{pt}$ 是 w_i 与实体 e_t 之间的距离向量. 连接 $\mathbf{w}_i^d$ 与 $\mathbf{w}_i^p$, 得到一个新的关于词 w_i 的向量 $\mathbf{w}_i = [\mathbf{w}_i^d; \mathbf{w}_i^p]$, 该向量的长度 $d_w = d + 2 \times d_p$, 用于 3.2.3 小节中的句子级语义特征提取.

3.2.2 语料级上下文特征抽取

CNSSNN 模型基于语料集的实体共现网络进行特征提取, 得到实体对的全局上下文特征. 这里的实体共现关系定义为任意两个实体在同一个句子中的共同出现关系. 我们从整个语料集的全部句子中

抽取所有的实体共现关系, 形成一个完整的实体共现网络. 该网络中的每一个实体, 都用 3.2.1 小节中描述的词嵌入向量 $\mathbf{w}^d$ 来表示. 与句子级局部上下文特征提取利用实体周围出现的词作为其上下文信息不同, 语料级全局上下文特征抽取使用实体在共现网络中的相邻节点来表示其上下文信息. 对于网络中任意实体节点 e_i , 设其相应的邻居集合 $N_i = \{e_{i1}, e_{i2}, \dots, e_{ij}\}$. 对一个实体而言, 其每个邻居节点在关系抽取任务中的重要程度都是不同的. 因此, 通过引入注意力机制, 为每个实体计算其各个相邻实体的不同权重, 然后用相邻实体的加权词向量和来表示该实体在共现网络中的上下文信息, 即得到该实体的语料级上下文特征 $\mathbf{f}_i^c$. 计算公式如下:

$$\mathbf{f}_i^c = \sum_{j=1}^{|N_i|} g_{ij} \times \mathbf{w}_j^d, \quad (1)$$

其中 $\mathbf{w}_j^d$ 为实体 e_i 的第 j 个相邻实体 e_{ij} 的 d 维词向量, g_{ij} 为该相邻实体对应的关系权重.

通过注意力机制为每个相邻实体 e_{ij} 计算其相对于实体 e_i 的关系重要程度, 计算公式如下:

$$a_{ij} = \tanh(\mathbf{W}_1[\mathbf{w}_j^d; \mathbf{w}_i^d] + b_1), \quad (2)$$

其中 $\mathbf{w}_i^d$ 为实体 e_i 的 d 维向量表示, $\mathbf{W}_1 \in \mathbb{R}^{1 \times 2d}$ 为参数矩阵, $b_1 \in \mathbb{R}^{1 \times 1}$ 为偏置. 在得到每个相邻实体相对实体 e_i 的关系重要程度分值 a_{ij} 之后, 对其进行归一化, 使得任意实体的所有相邻实体的权重和为 1, 从而得到各邻居实体的关系权重如下:

$$g_{ij} = \frac{\exp(a_{ij})}{\sum_{j=1}^{|N_i|} \exp(a_{ij})}. \quad (3)$$

我们通过注意力机制为每个实体 e_i 计算语料级特征向量 $\mathbf{f}_i^c \in \mathbb{R}^d$, 则句子中的每个实体对 $\langle e_h, e_t \rangle$ 将会得到两个语料级特征 $\mathbf{f}_h^c$ 和 $\mathbf{f}_t^c$. 连接两个特征向量, 将其经过一个线性全连接层, 并通过 ReLU 函数进行激活, 得到实体对的长度为 q 的语料级特征 $\mathbf{f}^c \in \mathbb{R}^q$.

$$\mathbf{f}^c = \text{ReLU}(\mathbf{W}_2[\mathbf{f}_h^c; \mathbf{f}_t^c] + b_2), \quad (4)$$

其中 $\mathbf{W}_2 \in \mathbb{R}^{q \times 2d}$ 为参数矩阵, $b_2 \in \mathbb{R}^q$ 为偏置.

3.2.3 句子级语义特征抽取

为了提取实体对 $\langle e_h, e_t \rangle$ 所在句子 s 的语义信息. 使用双向门控循环单元 (bi-GRU) 对输入语料进行隐语义提取. 假设当前样本为一个包含 n 个词的句子 $s = \{w_1, w_2, \dots, w_n\}$, 其中 w_i 代表句子中第 i 个词, 该词由 3.2.1 小节中描述的 d_w 维词向量 $\mathbf{w}_i$ 表示, 用 $\mathbf{S} \in \mathbb{R}^{n \times d_w}$ 表示一个句子的所有词向量组成的矩阵, 此时句子中的每个词互为相互独立的个体, 为了获取句子中词与词之间的语义连接, 将 $\mathbf{S}$ 输入 bi-GRU 中,

$$\vec{\mathbf{h}}_t = \overrightarrow{\text{GRU}}(\mathbf{w}_t, \vec{\mathbf{h}}_{t-1}), \quad (5)$$

$$\overleftarrow{\mathbf{h}}_t = \overleftarrow{\text{GRU}}(\mathbf{w}_t, \overleftarrow{\mathbf{h}}_{t+1}). \quad (6)$$

然后将 bi-GRU 的正向和反向隐藏层输出进行连接, 得到了句子的隐藏层输出 $\mathbf{h}_t$. 假设 bi-GRU 的隐藏层单元个数为 u , 则每个句子通过 bi-GRU 后的输出 $\mathbf{H}$ 为一个 $n \times 2u$ 的矩阵:

$$\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n), \quad (7)$$

其中 $\mathbf{h}_i = [\overleftarrow{\mathbf{h}}_i; \vec{\mathbf{h}}_i]$. 接着将 bi-GRU 的输出 $\mathbf{H}$ 通过一个最大池化层, 保留句子最重要的特征, 从而得到实体对 $\langle e_h, e_t \rangle$ 的句子级语义特征 $\mathbf{f}^s \in \mathbb{R}^{2u}$.

3.2.4 实体关系分类

将 3.2.2 小节描述的基于实体共现网络注意力机制提取的语料级上下文特征和 3.2.3 小节描述的基于 bi-GRU 提取的句子级语义特征联合起来, 采用分类的方法, 抽取实体关系.

使用 softmax 分类器为给定包含实体对 $\langle e_h, e_t \rangle$ 的句子 s 预测关系类型 $\hat{y}$. CNSSNN 模型将实体对的两部分特征, 即句子级语义特征 $\mathbf{f}^s$ 与语料级环境特征 $\mathbf{f}^c$, 输入一个全连接层, 进行特征融合, 再将融合后的特征 $\mathbf{f}$, 作为分类器的输入, 求出句子 s 中实体对 $\langle e_h, e_t \rangle$ 在关系标签 y 上的概率分布:

$$\mathbf{f} = \text{ReLU}(\mathbf{W}_3[\mathbf{f}^s; \mathbf{f}^c] + \mathbf{b}_3), \quad (8)$$

$$p(y|s) = \text{softmax}(\mathbf{W}_4 \times \mathbf{f} + \mathbf{b}_4), \quad (9)$$

其中 $\mathbf{W}_3 \in \mathbb{R}^{k \times (2u+q)}$ 和 $\mathbf{W}_4 \in \mathbb{R}^{r \times k}$ 为参数矩阵, $\mathbf{b}_3 \in \mathbb{R}^k$ 和 $\mathbf{b}_4 \in \mathbb{R}^r$ 为偏置, 超参 k 为融合后的特征向量大小, r 为分类标签数.

根据求得的标签概率分布, 可得实体对 $\langle e_h, e_t \rangle$ 的预测标签值为

$$\hat{y} = \arg \max p(y|s). \quad (10)$$

模型的损失函数定义为交叉熵损失函数

$$\mathcal{L} = - \sum_i \sum_j \text{label}_{ij} \log p(y_{ij}|s, \theta), \quad (11)$$

其中 θ 表示模型的所有参数, y_{ij} 表示样本 i 的预测标签为 j , label_i 为一个关于样本 i 的 one-hot 向量, 当其真实标签为 j 时 label_{ij} 为 1, 其余为 0.

4 实验

为了验证本文提出的 CNSSNN 模型的性能, 在两个数据集上将 CNSSNN 与其他现有的基于神经网络的关系抽取方法进行了对比, 验证了 CNSSNN 模型的性能.

4.1 数据集

4.1.1 CnNews

本数据集重点关注人与机构之间的关系抽取问题, 是一个人工标注数据集. 首先从互联网上抽取了 2017 年 3 月 23 日 ~ 11 月 2 日之间的 10 余万篇原始中文新闻文档, 然后将每篇文档拆分成单个的句子. 然后使用百度 NLP 分词工具对句子进行分词和命名实体识别, 将那些至少包含一个“人名”类, 命名实体和一个“机构”类命名实体的句子挑选出来.

接下来, 随机选取部分句子作为样本进行了人工标注, 将各句子中的实体对标注为: “持有” (hold)、“就读” (study at)、“就业” (work at) 和 “其他” (others) 共 4 类. 为了保证标注质量, 安排了 3 个人独立进行标注, 并按照多数投票原则决定最终的标签, 对于少量 3 个人都不一致的样本, 引入第 4 个人的意见进行裁决. 我们一共标注了 9040 个样本, 共包含 8613 个实体 (其中人名实体 3180 个, 机构名实体 5433 个). 标注数据集中各标签包含的样本数如表 2 所示.

为了构建实体的语料级上下文特征, 为上述 8613 个实体构建实体共现网络. 从剩余的未标注句子中将至少包含上述一个人名实体和一个机构名实体的句子挑选出来, 与标注的 9040 个句子一起, 最

表 2 标注数据集中各标签的样本数

Table 2 Numbers of samples of each label in the labeled dataset

Label	Number of samples
“hold”	1031
“study at”	923
“work at”	3033
“others”	4053
Total	9040

表 3 SemEval 中各标签的样本数

Table 3 Numbers of samples of each label in the SemEval dataset

Label	Number of samples
“others”	1864
“cause-effect”	1331
“instrument-agency”	660
“product-producer”	948
“content-container”	732
“entity-origin”	974
“entity-destination”	1137
“component-whole”	1253
“member-collection”	923
“message-topic”	895
Total	10717

终形成一个共包含了 96314 条边的实体共现网络. 该网络中节点的最大度数为 441, 最小度数为 1, 平均度数为 11.2.

在每轮实验中, 随机选取 70% 的样本作为训练集, 剩余 30% 的样本作为测试集.

4.1.2 SemEval

SemEval-2010 关系分类任务数据集^[15]包括“others”在内有 10 种关系类别, 共 10717 个样本, 其中 8000 个训练样本, 2717 个测试样本. 各类别标签包含的具体样本数如表 3 所示. 该语料所构建的实体共现关系网络共包含 6018 个节点, 9972 条边, 其中节点的最大度数为 118, 最小度数为 1, 平均度数为 3.3.

4.2 对比方法

将本文提出的 CNSSNN 模型与 4 种现有的基于神经网络的关系抽取方法进行了对比, 其中前两种是基于卷积神经网络的模型, 后两种是基于循环神经网络的模型.

(1) **CNN 模型.**经典的卷积神经网络模型. 我们使用词的 word2vec 向量来构造句子的表示矩阵作为 CNN 模型的特征输入, 然后将句子矩阵依次通过卷积和池化操作后, 输入 softmax 分类器, 求得样本在各类别标签上的概率大小.

表 4 模型中的超参设置
Table 4 Hyper-parameter setting of CNSSNN

Hyper-parameter	u	layer_num	q	batchsize	learning_rate	d	d _p
Value	100	1	64	250	1E-4	400	1

(2) **CR-CNN 模型**. 与经典的 CNN 模型相比, CR-CNN 模型^[8] 为句子中的每个词构建了两个额外的特征, 称之为词位置嵌入 (word position embedding, WPE), 这个特征表征了当前词与句子中心实体之间的相对距离.

(3) **bi-GRU 模型**. 门控循环单元 (GRU)^[25] 模型是循环神经网络 (RNN) 的一种改进. 与基于卷积和池化操作提取样本高层特征的 CNN 模型不同, 它通过加入门控机制对 RNN 模型进行优化, 能很好地处理语料中的远距离依赖问题. 本文采用双向 bi-GRU 作为对比模型之一.

(4) **Att-GRU 模型**. 为了更好地对注意力机制进行对比, 本文在 bi-GRU 的基础上, 加入句子级别的自注意力 (self-attention)^[26] 操作, 实现了一种 Att-GRU 模型. Att-GRU 在本质上与 Zhou 等^[23] 提出的 Att-BLSTM 是类似的, 都是基于双向循环神经网络, 加入注意力机制来提取出句子中各个词对当前关系类别的不同重要程度.

4.3 实验参数设置

本文提出的 CNSSNN 模型一共包含 5 个主要超参, 它们分别是: 门控循环单元的隐层单元个数 u ; 门控循环单元循环层数 layer_num ; 注意力机制全连接隐层单元个数 q ; 每次训练的样本数 batchsize 以及学习率. 本文在模型训练中使用 Adam^[27] 作为优化算法, 在语料级特征提取层、特征融合层与 softmax 分类层均加入 dropout 避免过拟合.

本文通过多轮实验对模型中的超参进行选择. 对于门控循环单元的隐层单元个数 u , 分别实验了 32, 64, 128, 256, 512 共 5 个参数值; 对门控循环单元循环层数 layer_num , 分别实验了 1, 2, 3 共 3 个参数值; 对注意力机制全连接层隐层单元个数 q , 分别实验了 32, 64, 128, 256 共 4 个参数值; 对 batchsize , 分别实验了 50, 100, 200, 400 共 4 个参数值; 对学习率 learning_rate 分别实验了 $1\text{E}-5$, $1\text{E}-4$, $1\text{E}-3$ 共 3 个参数值. 对于词向量维度 d , 则沿用了 Santos^[2] 论文中的设置. 对于用来表征词与实体相对距离的 WPE 向量的维度 d_p , 我们发现其设置对实验结果的影响不明显, 因此本文没有沿用文献 [8] 中的结论, 将其简单地设置为 1. 表 4 中列出了经过参数选择后模型所有超参的设置情况.

4.4 实验结果

本文用各种关系抽取方法在两个数据集上分别进行了 10 轮对比实验, 分别计算出每轮实验中各种方法的准确率、召回率及 $F1$ 值, 然后求平均值. 表 5 列出了全部标签的总体 $F1$ 值.

为了更好地比较不同的方法, 在抽取有意义的标签 (即无“其他”类) 关系时的性能表现, 表 6 列出了各种方法在排除“其他”标签之外的剩余标签上的准确率、召回率和 $F1$ 值.

从表 5 和 6 可以看出:

(1) 本文在关系抽取任务中, 基于循环神经网络的方法比基于卷积神经网络的方法有更好的整体性能. 因为循环神经网络能够克服卷积神经网络所不能兼顾到的远距离依赖问题, 可以更好地提取样本中前序和后序语义信息. 这说明, 在关系抽取任务中不能仅仅考虑实体及其临近的上下文信息, 还需要更远距离的语义信息和依赖关系.

表 5 各种关系抽取方法在全部标签上的性能比较

Table 5 Performance comparison of different relation extraction approaches on all labels

Model	<i>F1</i> on SemEval (%)	<i>F1</i> on CnNews (%)
CNN	80.43	85.32
CR-CNN	81.09	86.47
GRU	81.52	86.83
ATT-GRU	83.69	88.15
CNSSNN (ours)	85.99	90.34

表 6 各种关系抽取方法在排除“其他”标签上的性能比较

Table 6 Performance comparison of different relation extraction approaches without “other” label

Model	SemEval			CnNews		
	Precision (%)	Recall (%)	<i>F1</i> (%)	Precision (%)	Recall (%)	<i>F1</i> (%)
CNN	84.00	79.82	81.76	86.79	82.87	84.83
CR-CNN	84.20	80.82	82.40	86.85	85.99	85.86
GRU	82.85	81.07	81.94	85.91	86.16	86.21
ATT-GRU	85.19	83.07	84.06	87.21	87.93	87.58
CNSSNN (ours)	87.51	85.66	86.56	92.83	86.15	89.72

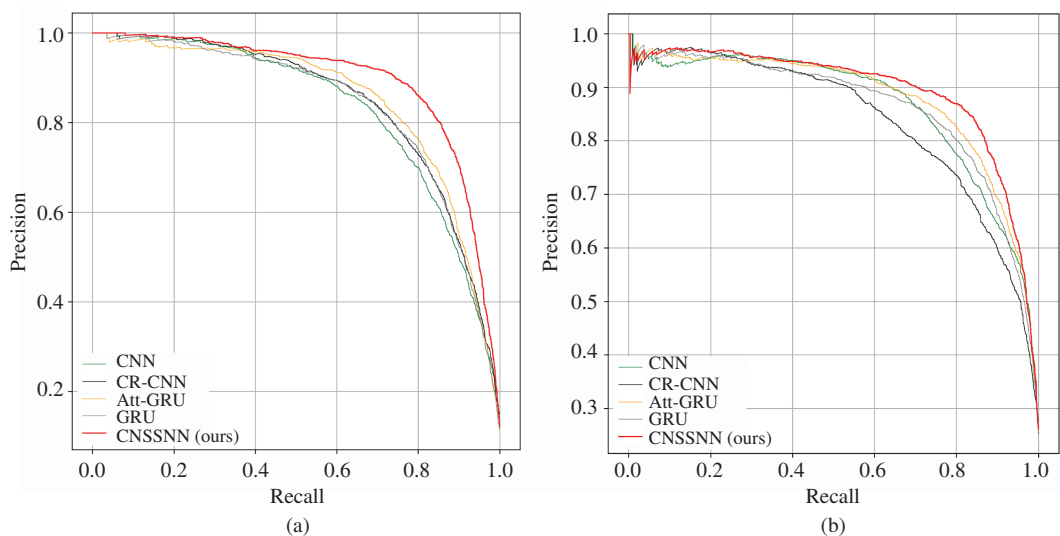


图 2 (网络版彩图) 各种模型的 P-R 曲线

Figure 2 (Color online) P-R curves of different approaches. (a) SemEval; (b) CnNews

(2) 在基于卷积神经网络的方法中, CR-CNN 模型比 CNN 模型有更好的表现, 这表明句子中词与实体的相对距离信息对实体关系抽取任务具有显著的提升作用.

(3) 在基于循环神经网络的方法中, Att-GRU 的性能要好于 GRU, 这表明句子中不同的词对关系抽取的贡献不同, 对各个词加以不同的权重能够提升关系抽取性能.

(4) 在全部方法的实验结果中, 本文所提出的 CNSSNN 模型比其他 4 种方法的表现都更好, 这充

分表明了 CNSSNN 模型的有效性. 其中 CNSSNN 模型与 GRU 模型的实验结果对比, 是语料级全局特征提升关系抽取性能的直接证明. 这也验证了本文所提出的观点: 实体的语料级全局上下文特征可以对句子级语义特征形成很好的补充, 从而有效地提升关系抽取的性能.

为了更直观地观察各种方法的性能表现, 图 2 展示了 CNSSNN 模型与其他 4 种对比方法的准确率-召回率 (precision-recall, P-R) 曲线对比情况. 从图 2 中可以直观地看出本文提出的 CNSSNN 模型能够在取得较高准确率的同时保证较高的召回率, 使得模型有很好的 $F1$ 值, 这也进一步验证了语料级的上下文特征对实体关系抽取有很好的辅助作用.

5 结论

本文针对当前实体关系抽取方法缺乏对语料级全局特征进行描述的问题, 提出了一种结合实体共现网络信息与句子语义信息的神经网络实体关系抽取模型 CNSSNN, 在利用 bi-GRU 提取句子级语义特征的同时, 通过在整个语料中构建实体共现网络, 并引入注意力机制为实体提取语料级全局上下文特征. 该模型能够同时捕获微观层面的句子级语义特征和宏观层面的语料级全局上下文特征, 克服了句子级特征可能存在的信息不充分的问题, 为实体关系抽取提供更丰富、更全面的分类信息. 为了验证提出的模型的有效性, 本文同时在两个实验数据集上进行了验证. 在两个数据集上的实验结果, 本文提出的 CNSSNN 模型与其他利用句子级语义信息进行关系抽取的方法相比具有更高的准确率与召回率, 验证了语料级全局特征对关系抽取任务的性能提升作用.

参考文献

- 1 Jurafsky D, Martin J. Speech and Language Processing. Beijing: Publishing House of Electronics Industry, 2018
- 2 Santos C N D, Xiang B, Zhou B W. Classifying relations by ranking with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015
- 3 Appelt D E, Bear J, Hobbs J R, et al. SRI international FASTUS system: MUC-4 test results and analysis. In: Proceedings of the 4th Conference on Message Understanding, 1992. 143-147
- 4 Yangarber R, Grishman R. NYU: description of the Proteus/PET system as used for MUC-7 ST. In: Proceedings of the 6th Message Understanding Conference, 1998
- 5 Zhou G D, Su J, Zhang J, et al. Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005
- 6 Bin X I, Qian L H, Zhou G D, et al. The application of combined linguistic features in semantic relation extraction. J Chinese Inf Process, 2008, 22: 44-49
- 7 Miao Q L, Zhang S, Zhang B, et al. Extracting and visualizing semantic relationships from Chinese biomedical text. In: Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation, 2012. 99-107
- 8 Zeng D J, Liu K, Lai S W, et al. Relation classification via convolutional deep neural network. In: Proceedings of the 25th International Conference on Computational Linguistics, 2014. 23-29
- 9 Wang L L, Cao Z, Melo G D, et al. Relation classification via multi-level attention CNNs. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 1298-1307
- 10 Ji G L, Liu K, He S Z, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017. 3060-3066
- 11 Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012. 1201-1211
- 12 Hashimoto K, Miwa M, Tsuruoka Y, et al. Simple customization of recursive neural networks for semantic relation classification. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2013. 1372-1376

- 13 Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 1105–1116
- 14 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. ArXiv:1409.0473
- 15 Hendrickx I, Su N K, Kozareva Z, et al. SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, 2009. 94–99
- 16 Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273–297
- 17 Jaynes E T. Information theory and statistical mechanics. Phys Rev, 1957, 106: 620–630
- 18 Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, 2001. 282–289
- 19 Zhang Y M, Zhou J F. A trainable method for extracting Chinese entity names and their relations. In: Proceedings of the 2nd Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, 2000. 66–72
- 20 Suchanek F M, Ifrim G, Weikum G. Combining linguistic and statistical analysis to extract relations from web documents. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006. 712–717
- 21 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst, 2013, 26: 3111–3119
- 22 Nanda K. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of ACL 2004 on Interactive Poster and Demonstration Sessions, 2013
- 23 Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 207–212
- 24 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013. ArXiv:1301.3781
- 25 Cho K, Merriënboer B V, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. ArXiv:1406.1078
- 26 Lin Z H, Feng M W, Santos C N D, et al. A structured self-attentive sentence embedding. In: Proceedings of International Conference on Learning Representations (ICLR), 2017
- 27 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations (ICLR), 2015

Combining entity co-occurrence information and sentence semantic features for relation extraction

Yudan MA, Yi ZHAO, Jing JIN & Huaiyu WAN*

Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

* Corresponding author. E-mail: hywan@bjtu.edu.cn

Abstract Relation extraction is one of the most important tasks in information extraction and a key step in knowledge graph construction. The existing relation extraction approaches mostly try to capture semantic features for entity pairs at the sentence level, which might ignore the global context information of the entities in the entire corpus. In this paper, we propose a novel neural network model for relation extraction, named CNSSNN, which combines the information of entity co-occurrences with sentences' semantic features. In this model, we first build an entity co-occurrence network from the corpus. Then, we introduce a network-level attention mechanism to capture network environmental information selectively and generate the corpus-level global context features for the entities. At the same time, we employ a bi-directional gated recurrent unit (bi-GRU) network to extract sentence-level semantic features for entity pairs. Finally, we combine the corpus-level features and the sentence-level features to classify relations. The experimental results, over a manually labeled dataset, show that our approach consistently outperforms other existing approaches in terms of both precision and recall.

Keywords information extraction, entity relation extraction, entity co-occurrence network, attention mechanism, gated recurrent unit



Yudan MA received a B.E. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2017. She is currently working towards a master's degree in the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include relation extraction and knowledge graph.



Yi ZHAO received a B.E. in biomedical engineering from Beijing Jiaotong University, Beijing, China, in 2017. He is currently working towards a master's degree in the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include knowledge graph and information extraction.



Jing JIN received a B.E. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2017. She is currently working towards a master's degree in the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include knowledge graph and knowledge representation learning.



Huaiyu WAN received his Ph.D. in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2012. He is an associate professor in the School of Computer and Information Technology, Beijing Jiaotong University. His research interests focus on social network mining, text mining, user behavior analysis, and traffic data mining.