基于依存关系和双通道卷积神经网络关系抽取方法

吴佳昌 吴观茂
(安徽理工大学计算机科学与工程学院 安徽 淮南 232001)

摘 要 关系抽取是自然语言中的一项重要任务,其结果对后续的信息抽取和自动问答系统有重要的影响。随着深度学习的日益火热,基于卷积神经网络的实体关系抽取已取得了不错的结果。不过词向量表示比较单一,提取的特征也有限。针对这个问题,将 Word2vec 训练的词向量和由自然语言处理工具得出的依存关系对分别作为模型两通道的输入向量,使用双通道卷积神经网络提取特征来实现实体关系抽取。该模型可以提取更深层的语义信息,并取得了比传统词向量更好的效果。

关键词 关系抽取 依存关系 卷积神经网络 双通道

中图分类号 TP183 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.04.038

RELATIONSHIP EXTRACTION METHOD BASED ON DEPENDENCY RELATION AND TWO-CHANNEL CONVOLUTIONAL NEURAL NETWORK

Wu Jiachang Wu Guanmao
(*School of Computer Engineering and Technology, Anhui University of Science and Technology, Huainan 232001, Anhui, China*)

Abstract Relation extraction is an important task in natural language, and its result has important influence on subsequent information extraction and automatic question-answering system. With the increasing popularity of deep learning, the entity relation extraction based on convolutional neural network has achieved good results. However, the word vector representation is relatively single, and the extracted features are limited. In view of this problem, the word vector trained by Word2vec and the dependency relationship obtained by the natural language processing tool were used as the input vectors of the two channels in the model. Two-channel convolutional neural network was adopted to extract features to realize entity relationship extraction. This model can extract deeper semantic information and achieve better results than traditional word vectors.

Keywords Relation extraction Dependency relationship Convolutional neural network Two-channel

0 引 言

关系抽取作为自然语言处理中最重要的任务之一,其结果直接影响着接下来信息抽取、机器翻译、自动问答系统等任务的进行,所以好的关系抽取结果则是进行其他任务的前提条件。

关系抽取方法主要分为三类:有监督、半监督和无监督的学习方法。有监督的学习方法虽然能够有不错的表现,能提取更有效的特征和更高的准确率和召回率,但比较依赖于人工标注语料。作为无监督的学习方法,其假设拥有相同语义关系的实体对拥有相似的上下文关系,用相似的上下文信息来对语义关系进行聚类,相对来说准确率没有有监督的学习方法高。

针对有监督学习方法需要大量人工标注语料和无监督学习方法的低准确率,Mintz 提出了远程监督的学习方法。这是一种标注方法,把知识库中存在的实体与实体的关系引入到正常的自然语言中进行训练。

作为深度学习中的重要技术,卷积神经网络有很强的提取深层特征的能力,在图片识别领域取得了重要的成功,在文本分类方面也表现不错。但是对于单通道、单一词向量的模型输入,所提取的特征还是不够

收稿日期:2018-10-19。国家自然科学基金项目(61471004)。**吴佳昌**,硕士生,主研领域:人工智能,数据挖掘。**吴观茂**,博士。

完整。本文重点研究有监督学习,并针对上述问题提出双通道卷积神经网络,一个通道输入词向量训练模型得出的句子向量,另一个通道输入句子的依存关系向量,然后进行特征的自动提取,通过 softmax 分类器进行分类。这个模型相对于单通道、单一词向量的卷积神经网络有一定的提升。

1 相关工作

关于实体关系抽取任务,经历了很长时间的探索和发展,也尝试采用了很多方法来实现关系抽取,比如说基于特征工程的方法[1]、基于核函数的方法等。有监督的分类方法是最常用的实现方法,而且也有很好的表现。近几年越来越多的人开始使用神经网络来解决关系抽取问题。Liu 等[2]第一次提出使用卷积神经网络(CNN)来解决关系抽取问题,虽然效果不错,但是因为结构比较简单、没有池化层,而存在较多噪声。文献[3]使用了较完整的 CNN 模型,pre-train 的词向量,而且其中加入了词的位置特征向量,不过只有一种卷积核,特征比较单一。针对卷积核大小单一的问题,文献[4]提出了多尺寸卷积核的 CNN 模型,不过还是传统的 CNN 结构,没有明显的提升。文献[5]在 CNN 结构上没有太多变化,只是采用了 Ranking Loss 的损失函数,比之前的 softmax 效果好。文献[6]采用双向 RNN(递归神经网络)结构,加入了实体的位置标签,取得的效果与 CNN 模型差不多。

文献[7]开始在大规模的数据集上做关系抽取,使用了多实例学习来减弱远程监督带来的噪声问题,而且在池化层也做了改进,采用了分段最大池化,更充分有效提取句子特征信息,也取得了较好的结果。文献[8]在文献[4]的基础上加入了 attention 机制,给赋予句子不同的权重,突出想要的信息,过滤掉噪声,充分利用信息,但是 CNN 模型还是适合用于短文本,RNN 模型更适合长文本。文献[9]中加入句子级别的 attention 机制,使用 CNN 模型结合静态词向量就取得了很好的效果。

从上述文献中可以看出,深度学习的方法在实体关系提取方面取得了不错的效果。如果加上的传统自然语言标注信息,效果可能更好。针对这个想法,本文提出了双通道卷积神经网络模型,其中利用自然语言处理工具加入了语义依存分析信息,能够深层次的提取两个单词之间的语义关系,通过实验证明了其有效性。

2 实体关系抽取

2.1 实体关系抽取模型

为了进一步提升深度学习实体关系抽取的能力,结合自然语言处理工具,将依存关系加入到特征提取中,使网络提取更深的提取能力。本文提出基于依存分析和双通道卷积神经网络的实体关系抽取模型如图 1 所示。此模型包含有映射层、卷积层、池化层和输出层。与一般卷积神经网络不同的是它拥有两个通道的输入向量:通道一使用 Word2vec 训练的词向量;通道二使用权威自然语言处理工具标注之后得到的词性和依存关系对向量。通过卷积核将两个通道的特征提取到一个总的特征图,再经过分段池化层的过滤,得到相对于最大池化更多的特征,最后送入分类器进行分类。

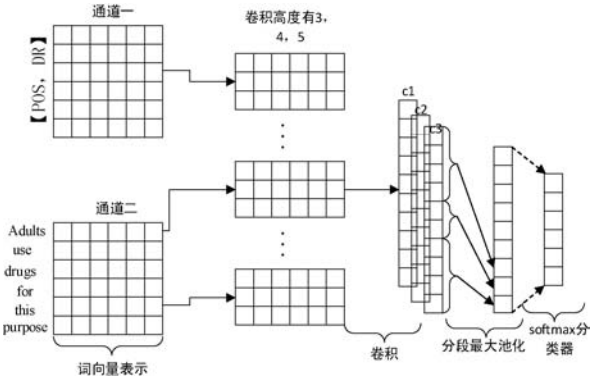


图 1 神经网络结构模型示意图

作为深度学习中一种著名的神经网络,卷积神经网络已经广泛应用于图像识别、语音识别等领域,并取得出色的结果。它有出色的特征提取能力,能够提取深层特征,并且还能消除一定的噪声。卷积神经网络包括输入层、卷积层、采样层和输出层。针对不同的任务,构造不同的模型结构,达到最好的效果。

整个实体关系抽取流程如图 2 所示。将数据集中的数据分别进行依存分析和词向量训练,得到的向量矩阵再送入双通道卷积神经网络中进行实体关系抽取。

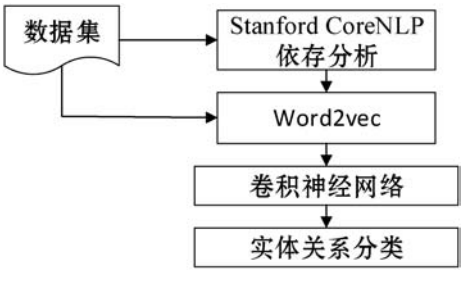


图 2 实体关系抽取流程图

2.2 词向量表示

词向量是词的一种基于神经网络的分布式表示,

是基于分布式假说对目标词进行建模,其思想就是相似的上下文信息,那么其所对应的语义关系也相似。主要的工作就是上下文的表示,以及对上下文和实体词之间的关系建模。词向量将每个词映射到一个某个 k 维的真值向量,维数较低而且能很好地捕获句子语义和句法特征,所以在很多 NLP 任务都有很好的表现。

目前对于训练词向量的方法常用的两种:Word2vec 和 GloVe。其中的 Word2vec 是由 Google 提供的开源词向量训练工序。CBOW 和 Skip-gram 是 Word2vec 中两种训练模型,CBOW 是通过上下文来预测目标词,而 Skip-gram 是通过目标词来预测上下文,总的思想都是相似的上下文有相似的目标词。本文使用 Skip-gram 来训练词向量,词向量的维数默认设置为 300 维。

为了更精确地表示和映射出词嵌入,我们在每个词嵌入的后面附加了其位置信息,就是体现当前这个词与句中实体的距离大小。其表达形式为(d1,d2)的一个元组。

例如某个单词的位置信息元组为(3, -2),3 为单词到实体 e1 的矢量距离,而 -2 为单词到实体 e2 的矢量距离,从左到右表示矢量正向,从右到左表示矢量负向,正数表示单词在实体的右侧,复数表示单词在实体的左侧。将每个词的位置信息加到每个词向量之后,组成一个新的向量。加入了位置信息的单词结构表现形式为【word,d1,d2】,其中:word 表示某个单词;d1 表示单词到实体 1 的距离;d2 表示单词到实体 2 的距离。词结构表现形式如表 1 所示。

表 1　词结构表现形式

word	d1	d2
Adults	0	-5
use	1	-4
drugs	2	-3
for	3	-2
this	4	-1
purpose	5	0

加入位置信息后的词序列按照单词的顺序划分为:

$$S = \{s_1, s_2, \cdots, s_n\}$$

然后使用 Word2vec 中训练好的 Skip-gram 模型得到词向量,每个词向量形如[0.792, -0.177, -0.107, 0.109, -0.542,…]是一种低维向量,相较于独热向量表示方法,既不会造成维数灾难也能表示词与词之间关联,包含的语义信息也更丰富。这些词向量按照词原有的顺序排列,组成一个矩阵,长度为 n,宽度为 d,然后将这个矩阵作为卷积神经网络一个通道的输入。

2.3　依存分析

依存分析的目的是为了发现两个单词之间的句法结构和语义关系,这种两个单词之间的关系被称为依存关系。

依存分析包括依存句法分析和语义依存分析。依存句法分析主要是通过识别句中词语成分如“主谓宾,定状补”,来分析各成分之间的依赖关系,并分析句子的句法结构。语义依存分析能够分析两个词语之间更深层次的语义关系,跨越句法结构的束缚,不同的句法结构可能表达的语义是相同的。文献[10]中使用 RNN 来得到句子的最短依存路径,然后加入到词向量中,使用卷积神经网络来提取特征,取得了不错的效果。

本文使用的是来自斯坦福大学开发的自然语言处理工具 Stanford CoreNLP,这是被广泛认可的自然语言处理工具。利用 Standford CoreNLP 对句子进行预处理,包括分词、词性标注、命名实体识别和依存句法分析。

依存关系对表现形式为:依存关系(依存词,核心词),依存关系是核心词和依存词之间的关系,而且这个关系是有方向的,依存词指向核心词,支配词支配依存词,这种支配关系不受距离的影响。

例如对句子“Adults use drugs for this purpose.”进行语义依存分析,结果如图 3 所示。图 3 中每个单词上方的大写字母缩写表示它们的词性,NNS 表示复数名词;VBP 表示非第三人称动词;IN 表示连词;DT 表示限定词;NN 表示单数名词。这个句子的依存关系对表示为:nsubj(use-2, Adults-1),root(ROOT-0, use-2),dobj(use-2, drugs-3),case(purpose-6, for-4),det(purpose-6, this-5),nmod(drugs-3, purpose-6)。nsubj 表示名词性主语,这里指的是“Adults”;root 表示根节点,一般指向句子中的谓语,这里指向“use”;dobj 表示直接宾语;case 表示格位标志;det 表示限定词;nmod 表示复合名词修饰。这些依存关系在图中表现为两个词通过一条有向弧线连接,由一端的依存词指向另一端的核心词。每个单词后所带的数字表示它在句子中的序号。

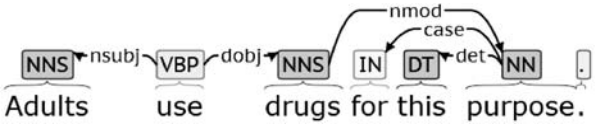


图 3　语义依存分析示例

本文首先将数据集中的数据进行依存分析之后,输出一个以单词为单位的元组,其结构表示为【POS,

DR】,其中:POS 表示句子中单词的词性;DR 表示依存关系。其对应的是每个单词和其依赖词汇组成的依存对,表现形式为 $r(n_1,n_2)$,r 是两个单词的依存关系;n_1 是依存词在句子中的序号;n_2 是核心词在句子中的序号。Stanford CoreNLP 输出结果如表 2 所示。这 n 个元组按词序排列,同样使用 Word2vec 训练得到词向量矩阵,这个矩阵宽度是 d,长度是 n(n 为句子的长度),记为 $R1 \in R^{n\times d}$,作为通道一的输入向量。

表 2 Stanford CoreNLP 输出结果

词性	依存关系	依存词序号	核心词序号
NNS	nsubj	2	1
VBP	root	0	2
NNS	dobj	2	3
IN	case	6	4
DT	det	6	5
NN	nmod	3	6

2.4 卷积过程

卷积层中,通过滑动卷积核(也称过滤器),与输入向量进行卷积操作,最后得到一个特征图。$\boldsymbol{R}1$,$\boldsymbol{R}2 \in \boldsymbol{R}^{d\times n}$ 作为卷积层的输入矩阵,其中 n 为句子 S 的长度(词的个数),d 为矩阵宽度。

两通道的输入矩阵宽度是一致的,从而方便了两个通道的卷积操作。两个向量矩阵上下排列,然后每个过滤器从通道一慢慢滑动到通道二,完成一次完整的卷积操作。

这样的连接方式好处是:1)减少了参数的个数,一定程度上减少了过拟合的可能性,从而提高了训练速度;2)破坏了结构对称性,从而能提取出不同特征。

过滤器的宽度为 d,与输入矩阵宽度一致,高度为 h,而过滤器的权值向量被随机初始化 W,其中 $W=\{w_1,w_2,\cdots,w_m\}$,有 m 个过滤器,每个过滤器包含 $h*d$个参数,$\boldsymbol{R}$ 表示输入向量矩阵,$\boldsymbol{R}[i:j]$表示矩阵 $\boldsymbol{R}$ 中第 i 行到第 j 行,其中 $j=i+h-1$。

在卷积过程中,有两种滑动情况如下:

1)过滤器在同一个通道滑动,即过滤器和矩阵 R 的第 i 行到第 j 行进行卷积操作。

$$c_{ki}=f(\boldsymbol{W}_k\cdot\boldsymbol{R}[i:j]+b)$$
$$1\leqslant i\leqslant n,\ i\leqslant j\leqslant n,\ j=i+h-1 \quad (1)$$

式中:b 表示偏置;f 为非线性激励函数 ReLU,能够快速得到较为精确的正确特征。

2)过滤器在两个通道中同时滑动。

$$c_{ki}=f(W_k\cdot\{\boldsymbol{R}1[i:n];\boldsymbol{R}2[1:j]\}+b)$$
$$n-h<i\leqslant n,i\leqslant j<h,\ j=h-n+i+1 \quad (2)$$

式中:$\boldsymbol{R}1$、$\boldsymbol{R}2$ 分别是通道一、通道二的输入向量矩阵。过滤器从通道一滑动到通道二,横跨两个通道,经过完整卷积操作之后得到一个特征图 C。

$$C=\{c_1,c_2,\cdots,c_m\}\in\boldsymbol{R}^{m\times(2s-h+1)}$$

式中:$2s-h+1$ 为特征图 C 高度;m 是 C 的宽度。

2.5 分段最大池化过程

池化层能降低输出结果的维度,还能保留最主要的特征。对于常见的最大池化操作,就是为了让输出的特征与其输入的句子长度无关,也就是无论输入的句子长度是多少,输出的特征长度是不变的。单独的最大池化操作被普遍使用,但是却不足以用于实体关系抽取。其用来提取特征太过粗糙,也无法准确提取两个实体之间的结构性信息,因此本文采用分段式最大池化操作。

如图 1 所示,将卷积层得到的特征图分为三段。第一段、第三段分别是对通道一和通道二进行卷积操作得到的特征图,第二段是对两通道同时进行卷积操作得到的特征图,将此特征图进行分段最大池化(piecewise max-pooling)。高度为 4、5 的过滤器也是同样操作。

$$p_{ki}=\max\{c_{ki}\} \quad (3)$$

式中:$1\leqslant k\leqslant n$,$i=1,2,3$ 对应着卷积层的三段输出,每一段进行最大池化操作得出 $p_k=\{p_{k1},p_{k2},p_{k3}\}$,将所有得到的 p_k 拼接起来,得到一个特征向量。记

$$p_{1:m}=[p_1,p_2,\cdots,p_m]$$

再通过一个非线性激励函数最后输出结果向量。

$$y=\tanh(p_{1:m}) \quad (4)$$

$y\in\boldsymbol{R}^m$ 的大小是固定的而且和输入句子的长度 n 无关。最后将得到特征向量送入 softmax 分类器进行分类。

2.6 Dropout 及分类层

针对深度学习训练过程中出现的过拟合问题,Hinton 提出了 Dropout 技术。过拟合问题有两个原因:一是训练样本数量太少,二是构建的模型复杂度过高。通常可以采用增加训练样本的数量、数据扩充和正则化约束。Dropout 思想就是正则化约束的一种实现形式,通过随机丢弃一些隐藏层神经元,使用修改过后的网络进行前向传播和反向传播,能够有效地防止过拟合的发生。

本文在最后第二层加入 Dropout 技术来实现正则化,在前向传播途中,以一定的概率 p 来丢弃一些隐藏层神经元来防止过拟合。

$$g=(y\otimes\boldsymbol{r}) \quad (5)$$

式中:$\boldsymbol{r}\in\boldsymbol{R}^m$ 是以概率为 p 的 Bernoulli 随机变量向

量; ⊗ 表示按位相乘; $y \in R^m$ 则是池化层的输出结果。最后将经过 Dropout 处理得到的向量 $\boldsymbol{g}$ 输入到 softmax 分类器中。

$$O = \boldsymbol{W}_t \boldsymbol{g} \tag{6}$$

式中: $\boldsymbol{W}_t \in \boldsymbol{R}^{m1 \times m}$ 是一个转换矩阵,而 $O = [O_1, O_2, \cdots, O_{m1}] \in \boldsymbol{R}^{m1}$, $\sum O_i = 1$,这是 softmax 层也是最后的输出结果。其中每个元素表示其对应类别的概率,概率最大元素所对应的类别就是分类结果。m1 表示可能的关系类型数目,实验数据使用的是 SemEval 2010 Task8 数据集,其中包含 10 种关系类型分类。

2.7　模型训练

本文模型中可训练参数为 $\theta = (\boldsymbol{R}1, \boldsymbol{R}2, W, \boldsymbol{W}_t)$,其中 $\boldsymbol{R}1$, $\boldsymbol{R}2$ 分别是通道一和通道二训练好的向量矩阵, $\boldsymbol{W} \in \boldsymbol{R}^{d \times h}$ 是卷积核参数, $\boldsymbol{W}_t$ 是 softmax 分类器的转换矩阵。本文使用的损失函数为:

$$J(\theta) = \sum_{i=1}^{N} \log p(o_i \mid x_i, \theta) \tag{7}$$

式中: N 表示样本的数量; o_i 表示输出的分类 i ; x_i 表示样本; θ 表示参数集合,其中条件概率 $p(o \mid x, \theta) = \frac{\exp(o_i)}{\sum_{k=1}^{m1} \exp(o_k)}$; $m1$ 是总的分类个数; o 是最后神经网络的输出结果。为了快速又准确的进行优化,本文采用 Adam 优化算法,进行快速地收敛损失函数,然后再使用随机梯度下降算法进行微调,使损失函数最小化。

3　实　验

首先介绍数据集,然后通过交叉验证确定模型参数,比较不同池化策略的池化效果,将本文方法和不加入词性和依存关系的模型在整体识别率和各个类别的识别率上相比较,最后得出结论依存关系对实体关系抽取的影响。

3.1　数据集

本文使用的数据集是 SemEval 2010 Task8 数据集,其中包括 8 000 条训练数据,2 717 条测试数据,数据集中训练样例如下所示:

"<e1>People</e1> have been moving back into <e2>downtown</e2>."

Entity-Destination(e1,e2)

Comment:

第一行是句子,其中两个实体已经通过"<e1>, </e1>, <e2>, </e2>"标注出来;第二行是两个实体的关系;第三行是备注。这个训练数据集包含 10 种实体关系,如表 3 所示。

表 3　10 种实体关系

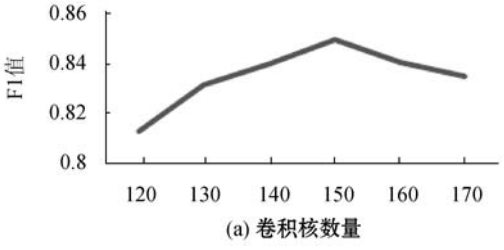
关系类型
Cause-Effect
Product-Producer
Entity-Origin
Instrument-Agency
Component-Whole
Content-Container
Entity-Destination
Member-Collection
Message-Topic
Other

3.2　超参数设置

表 4　超参数设置表

超参数	值
卷积核个数 m	150 * 3
卷积核高度 h	3,4,5
学习速率 η	0.01
Dropout 比率 D	0.6
词向量维度 d	300
正则化项	4
最小训练样本集值	50

其中词向量维度、学习速率、卷积核高度、正则化项设置为默认值,其他项通过十折交叉验证法得到。通过将训练集分成十份,轮流将其中一份作为测试集,其他 9 份作为训练集,经过多次实验得到上述参数。图 4 分别列出了卷积核个数和 Dropout 比率对 F1 值的影响。可以看出,在此实验中随着卷积核数量和 Dropout 比率的增加,F1 值逐渐增大,当卷积核数量为 150 和 Dropout 比率为 0.6 时,F1 值最高。但随着它们的继续增大,F1 值反而下降,说明过多的卷积核数量并不能提取更多有效特征,更高的 Dropout 比率反而会导致准确率的下降。



(a) 卷积核数量

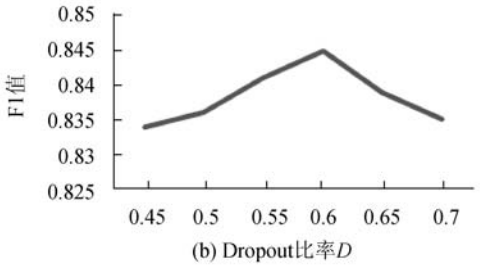


图 4 交叉验证结果

3.3 不同池化策略对比

不同的池化策略有不同的过滤效果,对最后的分类效果也会产生影响,表 5 讨论了在三种池化策略下模型的 F1 值表现。三段池化是本文采用的策略,根据卷积操作形成的三段特征图进行分段最大池化;二段池化则是以同时通过两通道的部分为界,对上下两部分进行最大池化。从结果看来,分段最大池化比最大池化效果有 0.8% 的提升,三段池化相对于二段池化有略微提升。

表 5 不同池化策略下的实验结果

池化策略	F1 值
最大池化	84.3
二段最大池化	84.9
三段最大池化	85.1

3.4 实验结果与分析

本文使用准确率、召回率和 F1 值来评价这个模型的性能。各模型性能参照如表 6 所示。

表 6 各模型性能表

模型	准确率	召回率	F1 值
CNNs	81.8	83.6	82.7
PCNNs	82.3	83.9	83.1
PCNNs + POS	83.7	84.1	83.9
PCNNs + DR	84.6	84.2	84.4
PCNNs + POS + DR	86	84.3	85.1

通过实验数据可以看出,在使用普通的 CNNs 和 PCNNs(分段卷积神经网络),F1 值只有 82.7 和 83.1,而 PCNNs 加上词性标注 POS,F1 值有 83.9;PCNNs 加上依存关系 DR 之后,F1 值变为 84.4;而本文模型(PCNNs + POS + DR)所获得的 F1 值为 85.1。本文方法相对于传统卷积神经网络来说有约 2.4% 的提升,也比单纯加入词性和依存关系的模型有提升。

另外,本文方法相对于 PCNNs 在每项类别上的识别效果如图 5 所示。使用本文方法后,相对于 PCNNs 有总体 2% 的提升,从每个类别识别效果来看,提升的大小不同,类别 Entity-Origin 和类别 Entity-Destination 大约有 3% 的提升,而类别 Instrument-Agency 和类别 Other 提升较少,说明加入了语义依存分析之后对整体识别效果有一定的提升,但是个别类别的识别效果提升不大。

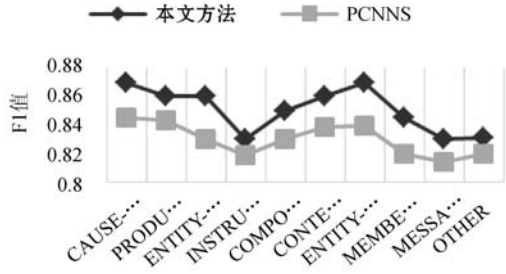


图 5 各项类别的 F1 值

根据实验可以得出以下结论 :

(1) 在引入依存分析作为特征之后,明显比单纯从 Word2vec 训练的词向量中提取特征来得更加准确,因为此模型能从语义层次提取句子中的信息,更好地反映句子的语法结构,分类性能也更好。综上,加入的特征越多,分类的结果也越准确,当然也得考虑不能有过多的参数,不然很难拟合。

(2) 使用卷积神经网络模型和自然语言处理工具相结合,比单一使用卷积神经网络等机器学习模型来得效果好,再加上卷积神经模型善于提取平面特征,能够出色地完成关系抽取任务。

4 结 语

本文针对单一使用训练之后的词向量提取特征或者自然语言处理工具来实现关系抽取,提出了一种基于依存关系的双通道卷积神经网络模型。Word2vec 训练的词向量和由自然语言处理工具得出的依存关系对分别作为模型两通道的输入向量,经过实验表明,两者结合能够有效提高 F1 值。不过这是基于有监督的情况下,更多的时候需要从一些无结构的语句中提出实体关系,而且要做到准确率高,还需要继续研究。

参 考 文 献

[1] 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.

[2] Liu C Y, Sun W B, Chao W H, et al. Convolution neural network for relation extraction[C]//International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013:231 - 242.

[3] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,2014: 2335 - 2344.

(下转第 267 页)

[2] Draisbach U, Naumann F. A Comparison and Generalization of Blocking and Windowing Algorithms for Duplicate Detection[C]//Proceedings of International Workshop on Quality in Databases, 2009:51-56.
[3] Bilenko M: Adaptive blocking: Learning to scale up record linkage [C]//International Conference on Data Mining. IEEE Computer Society, 2006:87-96.
[4] Hernández M A, Stolfo S J. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem[J]. Data Mining & Knowledge Discovery, 1998, 2(1):9-37.
[5] Papenbrock T, Heise A, Naumann F. Progressive Duplicate Detection[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(5):1316-1329.
[6] Ma M, Wang P, Chu C H, et al. Efficient Multipattern Event Processing Over High-Speed Train Data Streams[J]. IEEE Internet of Things Journal, 2015, 2(4):295-309.
[7] Yu Z, Kuang Z, Liu J, et al. Adaptive ensembling of semi-supervised clustering solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1577-1590.
[8] 郑津杨, 徐坤, 李建强. 用于 RFID 系统数据处理的排序邻居算法性能分析[J]. 计算机应用与软件, 2016, 33(12):207-210.
[9] 肖满生, 周浩慧, 王宏. 基于模糊综合评判的相似重复记录识别方法[J]. 计算机工程, 2010, 36(13):51-53.
[10] 刘雅思, 程力, 李晓. 基于长度过滤和动态容错的 SNM 改进算法[J]. 计算机应用研究, 2017, 34(1): 147-150.
[11] 杨巧巧, 郭振波, 王开西. 基于聚类分组和属性综合权值的 SNM 改进算法[J]. 工业控制计算机, 2017(9):27-28.
[12] Draisbach U, Naumann F, Szott S, et al. Adaptive Windows for Duplicate Detection[C]//IEEE International Conference on Data Engineering. IEEE, 2012:1073-1083.
[13] Christen P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 24(9): 1537-1555.
[14] Subramaniyaswamy V, Pandian S C. A Complete Survey of Duplicate Record Detection Using Data Mining Techniques [J]. Information Technology Journal,2012,11(8):941-945.
[15] 陈爽, 刁兴春, 宋金玉,等. 基于伸缩窗口和等级调整的 SNM 改进方法[J]. 计算机应用研究, 2013, 30(9):2736-2739.
[16] 邱越峰, 田增平, 季文赟,等. 一种高效的检测相似重复记录的方法[J]. 计算机学报, 2001, 24(1):69-77.
[17] 宋人杰, 余通, 陈宇红,等. 基于 MapReduce 模型的大数据相似重复记录检测算法[J]. 上海交通大学学报, 2018, 52(2):214-221.

(上接第 246 页)

[4] Nguyen T H, Grishman R. Relation Extraction: Perspective from Convolutional Neural Networks[C]//The Workshop on Vector Space Modeling for Natural Language Processing, 2015:39-48.
[5] Santos C N D, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks [J]. Computer Science, 2015, 86(86):132-137.
[6] Zhang D, Wang D. Relation classification via recurrent neural network[EB]. arXiv:1508.01006, 2015.
[7] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]//Conference on Empirical Methods in Natural Language Processing. 2015:1753-1762.
[8] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Meeting of the Association for Computational Linguistics,2016:2124-2133.
[9] Kim Y. Convolutional neural networks for sentence classification[C]//2017 XLIII Latin American Computer Conference (CLEI). IEEE, 2014: 1746-1751.
[10] Liu Y, Wei F, Li S, et al. A dependency-based neural network for relation classification[EB]. arXiv: 1507.04646, 2015.
[11] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,2014.

(上接第 254 页)

[3] 卢凯,徐建闽,叶瑞敏. 经典干道协调控制信号配时数解算法的改进[J]. 公路交通科技,2009,26(1): 120-124,129.
[4] 栗红强. 城市交通控制信号配时参数优化方法研究[D]. 长春:吉林大学,2004.
[5] Little J D C, Kelson M D, Gartner N M. MAXBAND: A program for setting signals on arteries and triangular networks [J]. Transportation Research Record, 1981, 795:40-46.
[6] Gazis D C. Traffic theory[M]. New York: Springer,2002.
[7] 何尚秋,郭海锋,俞立,等. 基于剪枝法的交通信号相位优化设计[C]//第三届国际电力电子与智能交通会议. 2010:425-428.
[8] 中华人民共和国公安部. GB/T 31418-2015 道路交通信号控制系统术语[S]. 北京:中国标准出版社,2015.
[9] Lu K, Zeng X, Li L, et al. Two-Way Bandwidth Maximization Model with Proration Impact Factor for Unbalanced Bandwidth Demands[J]. Journal of Transportation Engineering, 2012, 138(5):527-534.
[10] 王江静,姜久雷. 可伸缩矢量图形(SVG)[J]. 现代电子技术, 2005, 28 (24):32-33.