

文章编号: 1003-0077(2023)07-0032-10

基于语言模型增强的中文关系抽取方法

薛平¹, 李影^{1,2}, 吴中海^{1,2}

(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: 中文关系抽取任务旨在识别中文文本中实体对的语义关系。基于预训练语言模型的方法达到了当前最优结果。得益于大规模的训练语料, 预训练语言模型能够学习到训练语料中的语言知识, 提高了中文关系抽取模型的泛化能力, 但其庞大的参数规模需要消耗大量的存储和计算资源。为此, 该文提出了基于语言模型增强的中文关系抽取方法, 该方法基于多任务学习范式, 促进轻量级的中文关系抽取模型学习预训练语言模型中的语言知识。该文在三个中文关系抽取数据集上的实验结果表明了基于语言模型增强的中文关系抽取方法的有效性, 仅使用预训练语言模型 1% 的参数即可达到其 95% 的性能。

关键词: 中文关系抽取; 预训练语言模型; 语言模型增强

中图分类号: TP391

文献标识码: A

Language Model Enhanced Chinese Relation Extraction

XUE Ping¹, LI Ying^{1,2}, WU Zhonghai^{1,2}

(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: Chinese relation extraction task aims to extract the semantic relation between entity pairs in Chinese text. The pre-trained language model(PLL) based method has achieved best results nowadays. As an alternative for light-weight computing, we propose a language model enhanced Chinese relation extraction method, which learns the language knowledge in the pre-trained language model based on multi-task learning. This paper conducts experiments on three Chinese relation extraction datasets and the results that, using only 1% of the parameters of the pre-trained language model, the proposed method can achieve 95% of PLL performance.

Keywords: Chinese relation extraction; pre-trained language model; language model enhancement

0 引言

作为自然语言处理领域的重要任务之一, 关系抽取旨在识别文本中实体对之间的语义关系, 例如在句子“哈尔滨是黑龙江省会”中, 关系抽取方法可以提取出“哈尔滨”和“黑龙江”的语义关系是“位于”。抽取得到的实体关系三元组被用于知识图谱构建^[1]、智能问答系统^[2]、对话内容理解^[3]等下游任务, 并应用在多个领域, 包括医疗^[4]、金融^[5]、法律^[6]等。近年来随着深度学习方法的快速发展, 关系抽取任务取得了长足的进步, 大量基于深度学习的关

系抽取方法被提出。Liu 等人^[7]在 2013 年首次提出基于深度学习的关系抽取方法, 使用卷积神经网络对文本进行特征提取。之后, 陆续有研究者提出基于循环神经网络^[8-12]、基于图神经网络的方法^[13-14]等等。但上述关系抽取的研究工作, 主要集中于英文领域, 以单词序列作为输入。而在中文自然语言中, 单词与单词之间没有分隔符, 使用英文领域的关系抽取的方法需要先进行分词处理, 而分词过程中产生的错误, 会影响关系抽取任务的结果。为了避免错误传递, 目前大部分的中文关系抽取方法都是以字符序列作为输入。

基于字符序列的中文关系抽取方法, 虽然可以

收稿日期: 2021-12-31 定稿日期: 2022-04-05

基金项目: 国家重点研究与发展计划(2017YFB1002000)

避免分词处理所带来的错误传递问题,但却忽略了文本中单词所包含的高级别语义信息。例如“哈尔滨是黑龙江省会”中,“哈尔滨”“省会”等词所表示的城市的语义可以帮助推理出“位于”关系。目前,如何在中文字符序列中融合单词信息是中文关系抽取任务的主要研究内容之一,相关工作可以划分为基于词典的方法(lexicon-based)和基于预训练的方法(pretrain-based)。如图 1(a)所示,基于词典的方法通过将文本匹配固定的词典,获得文本中所有可能存在的单词,然后通过 Lattice-LSTM^[15] 或 Flat-Lattice^[16] 等网络将单词信息动态地编码进字符序列中;而基于预训练的方法,如图 1(b)所示,使用自然语言处理领域中流行的预训练语言模型,在大规模的语料中进行语言模型的训练。在训练的过程中通过全词遮蔽(Whole-Word-Mask)等方法^[17],将单词信息在预训练的过程中编码进模型,然后在中文关系抽取的任务上进行微调(Fine-Tune)。

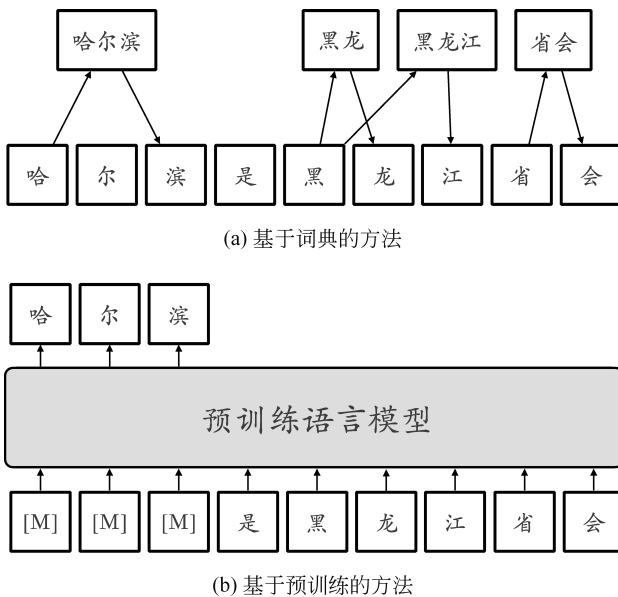


图 1 基于词典的方法和基于预训练的方法

预训练语言模型如 BERT^[18], 在大规模语料上进行遮蔽语言模型的训练。得益于复杂的模型结构和海量的训练数据,预训练语言模型可以对训练语料中的语言知识进行准确的刻画。相比于基于词典的方法,其包含的语言知识使得模型具有更好的泛化能力,能够抽取到更加准确的关系。但预训练语言模型的参数规模非常庞大,大部分的预训练语言模型都包含上亿的参数,在现实场景应用中需要消耗大量的存储资源和计算资源。

为了避免预训练语言模型高昂的模型开销,同

时利用其包含的语言知识,本文提出基于语言模型增强的中文关系抽取方法,其核心是采用多任务学习结构,在中文关系抽取模型基础上,训练一个语言模型来拟合预训练语言模型,学习预训练模型包含的语言知识。具体来说,我们使用轻量化的基于循环神经网络的关系抽取模型作为基准模型,同时以该循环神经网络训练一个双向语言模型。不同于一般的语言模型以下个字符作为标签训练,我们使用预训练语言模型的输出作为训练标签,使双向语言模型去拟合预训练语言模型,从而来学习其包含的语言知识。得益于多任务学习中的参数共享机制,双向语言模型学习到的语言知识可以对中文关系抽取模型进行增强,避免了预训练模型的高昂开销,同时有效利用预训练语言模型中的语言知识。

与我们方法类似的是基于知识蒸馏的模型压缩方法^[19],同样是使用轻量化的模型学习复杂模型的知识,但这两种方法的结构与所涉及的知识类型是不同的。知识蒸馏方法是使用复杂的老师模型在相关任务的数据集上先行进行学习,然后再将学习到的数据集内数据分布相关的知识蒸馏到轻量化的学生模型中,其中老师模型和学生模型都是关系抽取模型。而基于语言模型增强的中文关系抽取方法则是在关系抽取模型的基础上同时学习一个双向语言模型,通过双向语言模型来拟合预训练语言模型,来学习预训练模型中包含的语言知识。

本文的贡献主要包括:

(1) 针对基于预训练语言模型的中文关系抽取方法参数多及开销大的问题,提出了基于语言模型增强的中文关系抽取方法,采用多任务学习结构,训练双向语言模型从预训练语言模型学习语言知识,在高效抽取实体关系的同时避免了预训练模型的高昂开销。与知识蒸馏等压缩模型方法不同的是,本文提出的方法通过拟合预训练语言模型学习语言知识,突破了数据集的限制,提高了模型的泛化能力;

(2) 在三个中文关系抽取数据集上进行实验,证明本文提出的方法可以有效地提高基准模型的效果,超过了目前基于词典的方法。与基于预训练语言模型的方法相比,本文提出的方法使用其大约 1% 的参数量即可达到其 95% 的性能。

本文组织结构如下:第 1 节介绍相关工作;第 2 节给出中文关系抽取任务的定义和基准模型;第 3 节介绍基于语言模型增强的中文关系抽取方法;第 4 节给出实验结果和分析;第 5 节总结本文工作。

1 相关工作

目前,基于深度学习的关系抽取方法成为主流。Liu 等人^[7]在 2013 年首次探索了深度学习方法在关系抽取任务中的应用,提出基于卷积神经网络的关系抽取方法。之后,Zeng 等人^[20]在该方法的基础上加入了位置向量,用来表示文本中实体的位置信息。卷积神经网络的局限在于只能提取文本序列局部的特征信息。为了捕捉到全局的特征,在 2015 年,Zhang^[8]等人使用循环神经网络对文本和实体进行特征提取,以捕捉文本序列中的全局特征。在之后,为了进一步对文本进行表示,Zhang^[13]等人引入了额外的基于单词构建的依存句法树,使用图神经网络^[21]提取依存句法树的特征,来提高关系抽取任务的结果。

在中文关系抽取领域,为了避免分词错误导致的错误传递问题,主要工作都是字符级别的关系抽取方法,以字符序列作为输入。Zhang 等人^[22]指出了中文文本中分词错误问题导致的错误传递现象,提出了基于词典的方法并设计了在字符序列动态融合单词信息的 Lattice-LSTM 网络。之后,Li 等人^[15]提出了 MG-Lattice,在 Lattice-LSTM 的基础上融合了多粒度的信息,进一步提高中文关系抽取任务的准确率。Lattice-LSTM 虽然可以动态地融合单词信息,但是由于模型结构的限制,无法并行计算。为此,Zeng 等人^[16]使用可以并行计算的 Flat-Lattice 网路来动态地融合单词信息到字符序列中。

近些年,预训练语言模型 BERT^[18]被提出,大幅提高了大部分自然语言任务的指标。预训练语言模型在大规模的语料上进行遮蔽语言模型的学习,然后在特定的任务上进行微调(Fine-Tune)。得益于在大规模语料中学习到的语言知识,目前,大部分

的自然语言处理任务中最优的模型都是基于 BERT 或 BERT 的变体。同样地,基于预训练语言模型的中文关系抽取方法也得到了更优的结果。但预训练语言模型包含的参数非常庞大,大部分的预训练语言模型都包含上亿的参数,在现实世界的实际应用中,需要耗费大量的计算资源和存储资源。目前,针对预训练语言模型的压缩,有知识蒸馏^[23]、模型剪枝^[24]、模型量化^[25]等方法。与这些方法不同,本文提出的基于语言模型增强的中文关系抽取方法,直接学习预训练语言模型中包含的语言知识,使用这部分语言知识来增强中文关系抽取模型,相比于知识蒸馏等方法学习数据集相关的知识,语言知识能够突破数据集的限制,有效地提高模型的泛化能力。

2 中文关系抽取

本节介绍中文关系抽取的任务定义和中文关系抽取任务的基准模型。

2.1 任务定义

中文关系抽取一般被定义成分类问题,根据输入的文本和其中包含的实体对,对实体对之间的语义关系进行分类。将输入的文本表示为字符序列 $s = c_1, \dots, c_T$, 其中 T 表示字符序列的长度。字符序列中包含两个实体,将其表示为头实体 e^{head} 和尾实体 e^{tail} 。中文关系抽取任务的目的是根据字符序列 s , 提取出头实体 e^{head} 和尾实体 e^{tail} 之间的语义关系 $r \in R$, 其中 R 表示关系类型的集合。

2.2 基准模型

本文使用的基准模型是基于循环神经网络的中文关系抽取模型,如图 2 所示,包括嵌入层、编码层和分类层。

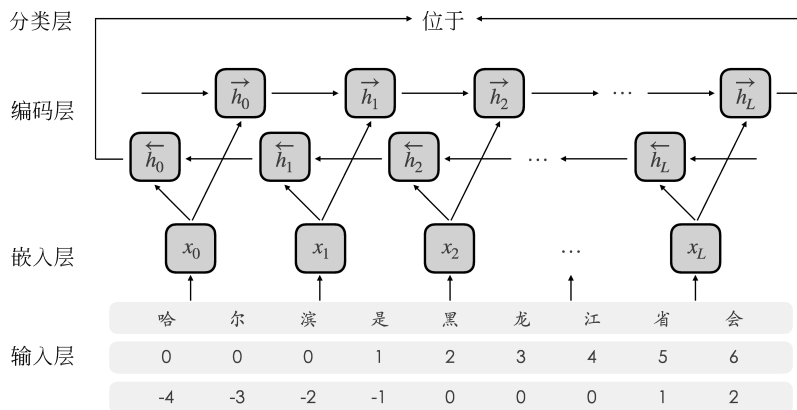


图 2 基于 LSTM 的中文关系抽取模型

嵌入层：首先,将字符序列映射到特征空间,得到每一个字符对应的向量表示。

$$\mathbf{x}_t^c = \text{CharEmbedding}(c_t) \quad (1)$$

其中, $\mathbf{x}_t^c \in \mathbb{R}^{d^c}$ 表示字符向量, d^c 表示字符向量的维度, 嵌入层使用预训练的字符向量初始化参数, 随训练过程进行更新。同时, 计算出每个字符与实体之间的相对位置, 相对位置信息可以根据文本中字符到实体之间的距离来帮助衡量不同字符对关系语义的贡献。相对位置的计算如式(2)所示, 以头实体为例, 计算每个字符与头实体之间的相对位置。

$$p_t^{\text{head}} = \begin{cases} t - b^h, & t < b^h, \\ 0, & b^h \leq t \leq e^h, \\ t - e^h, & t > e^h. \end{cases} \quad (2)$$

其中, p_t^{head} 表示每个字符与头实体的相对位置, b^h 和 e^h 分别表示头实体 e^{head} 首尾字符在字符序列中的索引位置, 每个字符与尾实体 e^{tail} 的相对位置计算同理。在获得每个字符与头实体和尾实体的相对位置后, 同样将其映射到特征空间, 获得相对位置的向量表示。

$$\mathbf{x}_t^{\text{head}} = \text{PositionEmbedding}_{\text{head}}(p_t^{\text{head}}) \quad (3)$$

$$\mathbf{x}_t^{\text{tail}} = \text{PositionEmbedding}_{\text{tail}}(p_t^{\text{tail}}) \quad (4)$$

其中, $\mathbf{x}_t^{\text{head}} \in \mathbb{R}^{d^p}$ 和 $\mathbf{x}_t^{\text{tail}} \in \mathbb{R}^{d^p}$ 分别表示每个字符相对头实体和尾实体的位置向量, d^p 表示相对位置向量的维度, 相关参数在训练开始时随机初始化。最后, 将字符向量和相对位置向量拼接, 获得嵌入层的最终表示 $\mathbf{x}_t$ 。

$$\mathbf{x}_t = [\mathbf{x}_t^c; \mathbf{x}_t^{\text{head}}; \mathbf{x}_t^{\text{tail}}] \quad (5)$$

编码层：编码层使用双向的 LSTM (Long Short-Term Memory)^[26] 网络作为编码器, 接收来自嵌入层的字符序列的特征表示, 计算全局的上下文相关特征。每个方向的 LSTM 网络根据上个时间步的输出特征和当前时间步的输入来计算当前时间步的输出特征。

$$\vec{h}_t = \overrightarrow{\text{LSTM}}(\mathbf{x}_t, \vec{h}_{t-1}) \quad (6)$$

$$\overleftarrow{h}_t = \overleftarrow{\text{LSTM}}(\mathbf{x}_t, \overleftarrow{h}_{t+1}) \quad (7)$$

其中 $\vec{h}_t \in \mathbb{R}^{d^h}$ 和 $\overleftarrow{h}_t \in \mathbb{R}^{d^h}$ 分别表示两个方向 LSTM 在时间步 t 的输出特征, d^h 表示 LSTM 的输出特征维度, 然后取两个方向 LSTM 最后一个时间步的结果作为编码层的最终表示。

$$\mathbf{h} = [\vec{h}_T; \overleftarrow{h}_0] \quad (8)$$

分类层：分类层在获得编码层的特征后, 首先将其映射到关系特征空间, 然后使用 Softmax 方法获得对应关系集合 R 中关系类别 r 的概率分布。

$$P(r | s, e^{\text{head}}, e^{\text{tail}}) = \text{Softmax}(W_o h + b_o) \quad (9)$$

其中, $P(r | s, e^{\text{head}}, e^{\text{tail}})$ 表示抽取得到的关系类别的概率分布, W_o 和 b_o 是分类层的训练参数。最后, 使用交叉熵函数计算中文关系抽取模型的目标函数 L_{re} 。

$$L_{\text{re}} = -\log(P(r | s, e^{\text{head}}, e^{\text{tail}})) \quad (10)$$

3 基于语言模型增强的中文关系抽取方法

本节介绍本文所提出的基于语言模型增强的中文关系抽取方法, 包括双向语言模型和语言模型增强两个部分。首先介绍在基准模型的结构之上构造双向语言模型, 其次介绍以预训练语言模型对双向语言模型进行增强。

3.1 双向语言模型

为了学习预训练语言模型中包含的语言知识, 首先在基于循环神经网络的基准关系抽取模型的基础上构造一个双向语言模型。

以前向语言模型为例, 根据前 $t-1$ 个字符来预测第 t 个字符的概率分布。基于多任务学习的结构, 双向语言模型和中文关系抽取模型共享嵌入层和编码层的参数。如图 3 左半部分所示, 语言模型

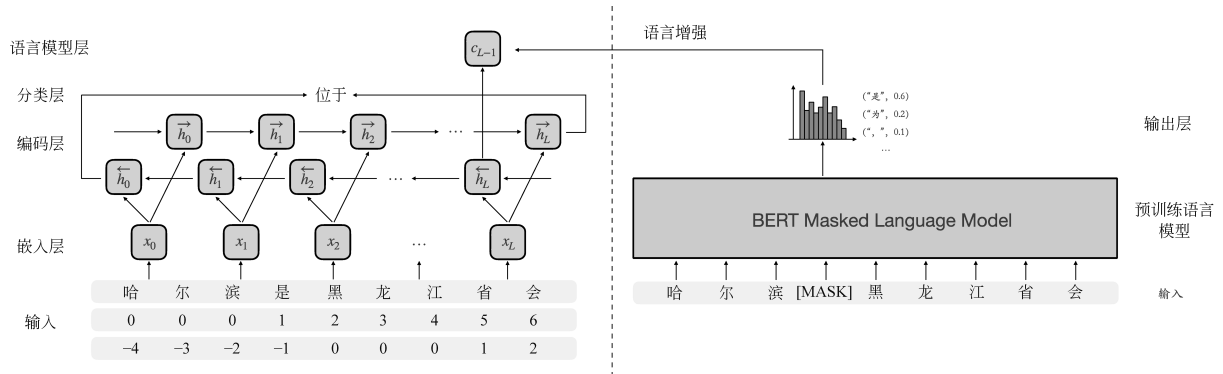


图 3 基于语言模型增强的中文关系抽取方法

层对 LSTM 网络的输出进行非线性特征映射,然后计算下个字符的概率分布。

$$\vec{m}_{t-1} = \tanh(\vec{W}_m \vec{h}_{t-1} + \vec{b}_m) \quad (11)$$

$$\overleftarrow{m}_{t+1} = \tanh(\overleftarrow{W}_m \overleftarrow{h}_{t+1} + \overleftarrow{b}_m) \quad (12)$$

$$P(c_t | \vec{m}_{t-1}) = \text{Softmax}(\vec{W}_q \vec{m}_{t-1} + \vec{b}_q) \quad (13)$$

$$P(c_t | \overleftarrow{m}_{t+1}) = \text{Softmax}(\overleftarrow{W}_q \overleftarrow{m}_{t+1} + \overleftarrow{b}_q) \quad (14)$$

其中, $P(c_t | \vec{m}_{t-1})$ 表示前向语言模型对第 t 个字符预测的概率分布,根据编码层在时间步 $t-1$ 的输出来预测第 t 个字符, $P(c_t | \overleftarrow{m}_{t+1})$ 则表示后向语言模型。 $\vec{W}_m, \vec{b}_m, \overleftarrow{W}_m, \overleftarrow{b}_m, \vec{W}_q, \vec{b}_q, \overleftarrow{W}_q, \overleftarrow{b}_q$ 分别表示双向语言模型中的训练参数。

3.2 语言模型增强

在中文关系抽取模型的基础上训练双向语言模型的目的是通过双向语言模型来拟合预训练语言模型,从而学习预训练语言模型中包含的语言知识。基于多任务学习的参数共享机制,通过双向语言模型学习到的语言知识可以对中文关系抽取模型进行增强。

学习预训练语言模型中包含的语言知识的本质,是使中文关系抽取模型学习更多样的关系表达,增加中文关系抽取模型的泛化能力。比如,当我们将“哈尔滨是黑龙江省会”中的“是”进行遮蔽,得到“哈尔滨[MASK]黑龙江省会”。然后使用预训练语言模型进行预测时,预测结果包含“是”“为”“,”等,用预测结果替换回遮蔽字符,可以得到“哈尔滨是黑龙江省会”“哈尔滨为黑龙江省会”和“哈尔滨,黑龙江省会”等多种组合。这些组合均表达了“哈尔滨”和“黑龙江”两个实体之间的“位于”关系,但是后两种表达是数据集内不包含的。基于语言模型增强的中文关系抽取方法可以看作是一种自动的数据增强(Data Augmentation)^[27]方法,通过预训练语言模型中的语言知识来对数据集进行数据增强,对每一条数据,生成多种不同但表达同样关系语义的文本,使中文关系抽取模型在训练过程中学习到多样的关系表达。

本文使用的预训练语言模型是 BERT-wwm^[28]。BERT-wwm 是一个 12 层深度的 Transformers 模型,在大规模的中文语料上进行遮蔽语言模型的训练,对输入文本进行部分字符的遮蔽,替换为固定的遮蔽符,通过被遮蔽字符的上下文信息来预测被遮蔽的字符。同时在训练的过程中,通过全词遮蔽(Whole-Word-Mask)的处理,在基于字符序列的预

训练语言模型的训练中融合单词信息。

对已经预训练好的 BERT-wwm 模型,输入包含遮蔽符的字符序列,可以得到被遮蔽的字符预测结果,表示为概率分布。如图 3 右半部分所示,对于字符序列“哈尔滨是黑龙江省会”中的“是”进行遮蔽,替换为遮蔽符“[MASK]”,输入到 BERT-wwm 模型中。通过 BERT-wwm 的预测,被遮蔽的字符可能为“是”“为”“,”等,括号内是其预测的概率值。BERT-wwm 对该遮蔽字符的预测结果是其包含的语言知识的具体表现,本质上是其对大规模训练语料的统计结果。我们将 BERT-wwm 对字符 c_t 的预测形式化为 $P_{\text{BERT}}(c_t)$ 。

$$P_{\text{BERT}}(c_t) = \text{BERT}(\dots, c_{t-1}, \text{mask}, c_{t+1}, \dots) \quad (15)$$

其中,mask 表示遮蔽符[MASK]。为了使双向语言模型拟合 BERT-wwm,如图 3 所示,将 BERT-wwm 对字符 c_t 的预测分布作为双向语言模型的训练标签。我们通过 KL 散度(Kullback-Leibler divergence)来拟合双向语言模型和预训练语言模型对字符预测的概率分布,作为双向语言模型的目标函数。

$$\vec{L}_{\text{lm}} = \frac{1}{T-1} \sum_1^T \text{KL}(P(c_t | \vec{m}_{t-1}) \| P_{\text{BERT}}(c_t)) \quad (16)$$

$$\overleftarrow{L}_{\text{lm}} = \frac{1}{T-1} \sum_0^{T-1} \text{KL}(P(c_t | \overleftarrow{m}_{t+1}) \| P_{\text{BERT}}(c_t)) \quad (17)$$

其中, $\vec{L}_{\text{lm}}$ 和 $\overleftarrow{L}_{\text{lm}}$ 分别表示前向语言模型和后向语言模型的目标函数。最后,我们将双向语言模型和中文关系抽取模型的目标函数进行联合优化,基于语言模型增强的中文关系抽取方法的整体目标函数为 L_{all} 。

$$L_{\text{all}} = L_{\text{re}} + \lambda(\vec{L}_{\text{lm}} + \overleftarrow{L}_{\text{lm}}) \quad (18)$$

其中, λ 表示双向语言模型目标函数的权重,通过调整 λ 来平衡多任务学习结构中两部分模型的训练过程。

4 实验分析

本节通过实验验证证明基于语言模型增强的中文关系抽取方法的效果。首先在 4.1 节中描述实验中使用的数据集和评估指标。在 4.2 节对实验环境设置进行描述,包括方法中使用的工具和重要的超参数。然后在 4.3 节中,与多种基线方法进行多维度的比较。最后在 4.4 节,对本文提出的方法进行消融实验。

4.1 实验数据与评价指标

本文使用三个中文关系抽取数据集作为实验数据,分别是 SanWen^①、ACE2005^② 和 ACE2004^③。其中,SanWen 是根据文学数据构建的,一共包含 10 类关系,并按照一定比例被划分为训练集、验证集和测试集。ACE2005 和 ACE2004 数据集是新闻领域的关系抽取数据集,分别包含 18 类和 19 类关系,对于 ACE2005 和 ACE2004 数据集的划分,我们和 Li 等人^[15]的研究工作保持一致,我们随机选取其 75% 的数据训练,使用剩下的 25% 的数据进行验证。数据集的统计信息如表 1 所示。

表 1 实验数据集统计

数据集	样本数	平均长度	关系数
SanWen	21 240	30.0	10
ACE2005	7 128	27.4	18
ACE2004	5 842	28.3	19

为衡量不同中文关系抽取方法在上述数据集上的效果,我们使用分类任务中的 F_1 (%) 值和 AUC (%) 值作为评价指标。

4.2 实验设置

本文使用 PyTorch 深度学习库来实现本文的方法,为了与相关工作进行合理的比较,实验环境中涉及的大部分超参数都和 Li 等人^[15]研究工作中使用的保持一致。我们使用 Li 等人^[15]工作中预训练的 100 维字符向量对嵌入层参数进行初始化,相对位置向量的维度设为 10 维,在训练开始时进行随机初始化,两者在训练过程中进行更新。编码层使用双向 LSTM,每个方向的 LSTM 的隐藏层维度设为 200。同时,我们对嵌入层和编码器的输出特征分别进行 Dropout,比例为 0.5。

在语言模型增强的部分,我们使用在中文维基数据上进行预训练的 BERT-wwm 模型。预训练语言模型的输出结果是所有字符上的概率预测,其中只有部分字符是有意义的预测结果,其余则没有意义。为了减少无意义字符对结果的影响,需要对预训练语言模型的预测结果进行截断,每个字符的预测结果按输出值的大小取前 100 个预测结果,根据截断后的输出值重新计算这 100 个字符的概率分布,作为双向语言模型的标签。

对于模型的训练,我们使用 Adam 优化器^[29],学

习率设为 0.000 5。在训练过程中,双向语言模型目标函数的权重 λ 设为 0.01。重要超参数的值如表 2 所示。

表 2 实验相关超参数

参数	取值
Epoch	100
Batch Size	32
字符向量维度	100
相对位置向量维度	10
LSTM 隐藏层维度	200
优化器	Adam
学习率	0.000 5
dropout	0.5
λ	0.01

4.3 实验结果

本节将基于语言模型增强的中文关系抽取方法与多个基线模型进行对比实验,证明基于语言模型增强的方法能够有效降低预训练语言模型的开销。在 4.3.1 节,与非预训练模型进行对比;在 4.3.2 节,与预训练模型在关系抽取任务指标和模型参数量上进行对比。

4.3.1 与非预训练模型的对比结果

首先,与非预训练的模型做对比实验,首先实现了两种经典的关系抽取模型,分别是 Zeng 等人^[20]提出的基于 CNN 的关系抽取模型和 Zhang 等人^[9]提出的基于 LSTM 的关系抽取模型。同时,我们也和两种目前流行的基于词典的中文关系抽取方法进行对比,分别是 Li 等人^[15]提出的 MG-Lattice 模型和 Zeng 等人^[16]提出的 Flat-Lattice 模型。其中,考虑到 ACE2005 数据集随机划分可能造成的数据分布不一致,我们使用 MG-Lattice 方法公开的源代码进行实验,获得实验结果。Flat-Lattice 方法则使用其论文中给出的结果进行对比。与非预训练模型的对比结果如表 3 所示。

(1) 首先,我们可以观察到,基于语言模型增强的中文关系抽取方法,在表中表示为“LSTM+语言

① <https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset>

② <https://catalog.ldc.upenn.edu/LDC2006T06>

③ <https://catalog.ldc.upenn.edu/LDC2005T09>

表 3 与非预训练模型的对比结果

(单位: %)

模 型	SanWen		ACE2005		ACE2004	
	F_1	AUC	F_1	AUC	F_1	AUC
CNN	64.67	56.40	80.46	76.34	78.60	74.87
LSTM	64.68	57.60	79.73	74.99	77.08	72.25
MG-Lattice	65.61	57.33	80.05	75.36	79.21	75.46
Flat-Lattice	68.35	59.14	79.81	74.31	/	/
LSTM+语言模型增强	70.88	65.88	84.07	79.94	81.69	78.35

模型增强”,在三个数据集的 F_1 值和 AUC 值均达到了最高。在 SanWen 数据集上,与目前最好的基于词典的中文关系抽取模型 Flat-Lattice 相比, F_1 值提高了 2.53%,AUC 值提高了 6.74%,在其他数据集上也有相应的提升。

(2) LSTM+语言模型增强与 LSTM 模型相比,仅增加了以预训练模型的预测结果作为训练标签的双向语言模型。对比 LSTM 模型,LSTM+语言模型增强在三个数据集上的提升即是语言模型增强的效果。在 SanWen 数据集上, F_1 值提升了 6.2%,AUC 值提升了 8.28%。证明了预训练语言模型的语言知识确实能够对中文关系抽取模型进行增强。

(3) 最后,我们通过模型的预测结果进行示例

分析,我们选择 LSTM 方法和 MG-Lattice 方法与本文所提出的基于语言模型增强的方法进行对比,示例结果如表 4 所示。我们在 SanWen 数据集测试集内选择了两条文本作为示例,从预测结果中可以看出,LSTM 方法与 MG-Lattice 方法预测出了错误的关系类型,而我们所提出的方法可以预测出正确的关系类型。对于第二个示例,“当包谷杆上长出第一个包谷棒时”这条文本中,如果实体 1 是“包谷杆”,与实体 2“包谷棒”的关系则是 LSTM 方法与 MG-Lattice 方法预测的“部分整体”。但实体 1 是“包谷杆上”,与实体 2“包谷棒”的关系是“位于”。该示例结果证明本文所提出的基于语言模型增强的方法可以准确地根据实体的细微差别识别出正确的关系类型,具有更强的关系抽取能力。

表 4 模型预测结果对比

文 本	实 体 1	实 体 2	关 系	预测结果		
				LSTM	MG-Lattice	LSTM+语言模型增强
梁一元赶紧用云南白药止血贴	梁一元	云南白药	使用(Use)	位于(Located)	位于(Located)	使用(Use)
当包谷杆上长出第一个包谷棒时	包谷杆上	包谷棒	位于(Located)	部分整体(Part-Whole)	部分整体(Part-Whole)	位于(Located)

4.3.2 与预训练模型的对比结果

为了与基于预训练语言模型的方法进行对比,我们实现了 Livio 等人^[30]所提出的基于预训练语言模型的关系抽取方法,使用 BERT-wwm 预训练模型。同时,为了比较不同参数规模下预训练语言模型的效果,我们分别取 BERT-wwm 的前 6 层的参

数和前 3 层的参数进行实验。同时,我们也与 Tang 等人^[23]提出的针对预训练语言模型的知识蒸馏方法进行对比。除了比较三个数据集上的分类指标,同时也比较模型的参数规模,表中表示为 Parameter,含义为模型的参数量。与预训练模型的对比结果如表 5 所示。

表 5 与预训练模型的对比结果

(单位: %)

模 型	SanWen		ACE2005		ACE2004		Parameter
	F_1	AUC	F_1	AUC	F_1	AUC	
BERT-wwm	76.18 (100%)	73.71 (100%)	88.34 (100%)	85.63 (100%)	85.75 (100%)	82.93 (100%)	102.27M (100%)

模 型	SanWen		ACE2005		ACE2004		Parameter
	F_1	AUC	F_1	AUC	F_1	AUC	
BERT-wwm (6 层)	72.86 (95.6%)	68.49 (92.9%)	85.70 (97.0%)	82.50 (96.3%)	83.21 (97.0%)	80.41 (96.9%)	59.74M (58.4%)
BERT-wwm (3 层)	69.90 (91.7%)	64.41 (87.3%)	83.39 (94.4%)	79.43 (92.7%)	81.62 (95.1%)	77.96 (94.0%)	38.48M (37.6%)
LSTM+知识蒸馏	67.60 (88.7%)	59.53 (80.7%)	81.64 (92.4%)	76.65 (89.5%)	80.17 (93.4%)	73.78 (88.9%)	0.81M (0.8%)
LSTM+语言模型增强	70.88 (93.0%)	65.88 (89.4%)	84.07 (95.2%)	79.94 (93.4%)	81.69 (95.3%)	78.35 (94.5%)	1.29M (1.3%)

(1) 我们可以从表中观察到,BERT-wwm 在三个数据集上均达到最好,LSTM+语言模型增强与BERT-wwm 模型在 F_1 值和 AUC 值上仍然有差距。但 BERT-wwm 的参数规模非常庞大,达到了 102.27M(M 表示百万)。我们以 BERT-wwm 作为基准(表示为 100%),计算其他方法与其在性能和参数规模上的百分比。本文提出的基于语言模型增强的方法在三个数据集上的平均性能都可以达到 BERT-wwm 的 95%,而参数规模仅为 1.29M,为 BERT-wwm 的 1.3%。实验证明了基于语言模型增强的中文关系抽取方法可以在达到相似性能的同时大幅降低预训练模型的开销。

(2) BERT-wwm (6 层)和 BERT-wwm (3 层)分别仅使用 BERT-wwm 的部分参数。可以观察到,随着层数的降低,基于预训练语言模型的方法在三个数据集上的指标也随之降低。在使用 BERT-wwm (3 层)进行实验时,LSTM+语言模型增强的方法在三个数据集上的指标上已经超过 BERT-wwm (3 层)方法。但 BERT-wwm (3 层)的参数规模仍然是 LSTM+语言模型增强方法的 30 倍。证

明相比于使用小规模预训练语言模型,基于语言模型增强的中文关系抽取方法是一个更优的选择。

(3) 最后,我们与基于知识蒸馏的方法(表中表示为 LSTM+知识蒸馏)相比,我们的方法在三个数据集上的指标均超过基于知识蒸馏的方法。实验证明了相比于数据集内的知识,语言知识对于中文关系抽取任务的提升更大。

4.4 消融实验

本节对基于语言模型增强的中文关系抽取方法进行消融实验,对方法中的模块进行消融,分析各个模块带来的性能提升。消融实验的结果如表 6 所示。首先,我们对来自预训练语言模型的增强进行剥离,仅进行标准的双向语言模型训练,以下个字符作为标签,在表中表示为 LSTM+语言模型。可以观察到,在剥离了预训练语言模型的预测结果时,三个数据集的指标均有所下降,在 SanWen 数据集上, F_1 值下降 3.21%,AUC 值下降 3.74%。下降的结果即是来自预训练语言模型的语言知识对中文关系抽取任务的增强结果。

表 6 消融实验结果							(单位: %)
模 型	SanWen		ACE2005		ACE2004		
	F_1	AUC	F_1	AUC	F_1	AUC	
LSTM+语言模型增强	70.88	65.88	84.07	79.94	81.69	78.35	
LSTM+语言模型	67.67	62.14	82.32	78.19	80.11	75.76	
LSTM	64.68	57.60	79.73	74.99	77.08	72.25	

之后,我们继续剥离掉方法中基于多任务学习结构中的语言模型的部分,方法退化为标准的基于 LSTM 的中文关系抽取方法。三个数据集的指标继续下降,下降的结果为基于多任务学习的语言模

型带来的提升,在 SanWen 数据集上, F_1 值下降 2.99%,在 AUC 值上下降 4.54%。在消融实验中,我们分别对基于语言模型增强的中文关系抽取方法中的两个核心模块进行消融,验证了每个模块对中

文关系抽取任务带来的提升效果。

5 总结

针对预训练模型参数量多开销大的问题,本文提出基于语言模型增强的中文关系抽取任务,采用多任务学习结构,在中文关系抽取模型的基础上训练双向语言模型,以训练语言模型的预测结果作为训练标签,通过拟合预训练语言模型来学习预训练语言模型中包含的语言知识。预训练语言模型中的语言知识可以对训练数据进行自动的数据增强,使中文关系抽取模型在训练过程中学习到多样化的关系表达,提高中文关系抽取方法的性能和泛化能力。

未来计划对预训练语言模型中的语言知识进行更细粒度的蒸馏,进一步提升语言模型增强方法的性能。

参考文献

- [1] TRISEDYA B D, WEIKUM G, QI J, et al. Neural relation extraction for knowledge base enrichment[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy; Association for Computational Linguistics, 2019: 229-240.
- [2] DENG Y, XIE Y, LI Y, et al. Multi-task learning with multi-view attention for answer selection and knowledge base question answering[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, 2019: 6318-6325.
- [3] LIU J, SUI D, LIU K, et al. Graph-based knowledge integration for question answering over dialogue[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 2425-2435.
- [4] XING R, LUO J, SONG T. BioRel: Towards large-scale biomedical relation extraction[J]. BMC Bioinform, 2020, 21-S(16): 543.
- [5] ZHOU Z, ZHANG H. Research on entity relationship extraction in financial and economic field based on deep learning[C]//Proceedings of the IEEE 4th International Conference on Computer and Communications, 2018: 2430-2435.
- [6] CHEN Y, SUN Y, YANG Z, et al. Joint entity and relation extraction for legal documents with legal feature enhancement[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 1561-1571.
- [7] LIU C, SUN W, CHAO W, et al. Convolution neural network for relation extraction[C]//Proceedings of Advanced Data Mining and Applications-9th International Conference, 2013: 231-242.
- [8] ZHANG D, WANG D. Relation classification via recurrent neuralnetwork[J]. arXiv preprint arXiv:1508.01006, 2015.
- [9] MIWA M, BANSAL M. End-to-end relation extraction using lstms on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1105-1116.
- [10] XIAO M, LIU C. Semantic relation classification via hierarchical recurrent neural network with attention[C]//Proceedings of the 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 2016: 1254-1263.
- [11] YU B, ZHANG Z, LIU T, et al. Beyond word attention: using segment attention in neural relation extraction[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 5401-5407.
- [12] LUO X, ZHOU W, WANG W, et al. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geologicaldata[J]. IEEE Access, 2018, 6: 5705-5715.
- [13] ZHANG Y, QI P, MANNING C D. Graph convolution over pruned dependency trees improves relation extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 2205-2215.
- [14] GUO Z, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 241-251.
- [15] LI Z, DING N, LIU Z, et al. Chinese relation extraction with multi-grained information and external linguistic knowledge[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 4377-4386.
- [16] ZENG X, ZHONG J, WANG C, et al. Chinese relation extraction with flat-lattice encoding and pretrain-transfer strategy[C]//Proceedings of Knowledge Science, Engineering and Management-14th International Conference, KSEM, 2021: 30-40.
- [17] SUN Y, WANG S, LI Y-K, et al. ERNIE: enhanced representation through knowledgeintegration[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1441-1451.

- [18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [19] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
- [20] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics, Proceedings of the Conference; Technical Papers, 2014: 2335-2344.
- [21] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the 5th International Conference on Learning Representations, 2017: 24-26.
- [22] ZHANG Y, YANG J. Chinese ner using lattice lstm [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.
- [23] TANG R, LU Y, LIU L, et al. Distilling task-specific knowledge from bert into simple neural networks [J]. arXiv preprint arXiv, 1903.12136, 2019.
- [24] ZAFRIR O, LAREY A, BOUDOUKH G, et al. Prune once for all: Sparse pre-trained languagemodels [C]//Proceedings of Near IPS, 2021.
- [25] BAI H, HOU L, SHANG L, et al. Towards efficient post-training quantization of pre-trained languagemodels [J]. arXiv preprint arXiv, 2109.15082, 2021.
- [26] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [27] FENG S Y, GANGAL V, WEI J, et al. A survey of data augmentation approaches for nlp[C]//Proceedings of the Association for Computational Linguistics; ACL/IJCNLP, 2021: 968-988.
- [28] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese bert[J]. arXiv preprint arXiv, 1906.08101, 2019.
- [29] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations, 2015: 1-15.
- [30] SOARES L B, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 2895-2905.



薛平(1993—), 博士研究生, 主要研究领域为关系抽取。

E-mail: xue.ping@pku.edu.cn



吴中海(1968—), 博士, 教授, 主要研究领域为大数据系统与分析、大数据与云安全、嵌入式系统。

E-mail: wuzh@pku.edu.cn



李影(1975—), 通信作者, 博士, 教授, 主要研究领域为分布式计算、可信计算。

E-mail: li.ying@pku.edu.cn