doi:10.3772/j.issn.2095-915x.2021.06.004

融合多头自注意力的远程监督关系抽取方法

李思琪 朱庆 陈钰枫 徐金安 张玉洁

北京交通大学计算机与信息技术学院 北京 100044

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 在关系抽取的研究领域中，通过远程监督方法可以快速地获取大量标注好的训练数据，但是其存在的关系标签错误标注问题会引入大量噪声数据。已有的研究工作主要使用注意力机制来降低噪声数据的影响，但这种方法在注意力分配时使用的是含有噪声的关系标签，可能导致“越学越错”的问题。[方法 / 过程] 本文提出了一种融合多头自注意力的远程监督关系抽取方法，在词级别注意力分配上，使用多头自注意力计算并分配权重，在句子级别注意力分配上，使用头、尾两实体的语义表征作为权重分配的依据，从而避免采用有噪声的关系标签作为注意力分配的依据，以降低噪声的影响。[结果 / 结论] 在公开数据集上的实验结果表明，相较于使用有噪声的关系标签来分配注意力，所提方法的性能有了显著提高。

关键词：关系抽取；远程监督；多头自注意力；实体特征

中图分类号：G35；TP391

Distant Supervision for Relation Extraction with Multi Head Self Attention

LI Siqi ZHU Qing CHEN Yufeng XU Jin’an ZHANG Yujie

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

基金项目 国家自然科学基金“面上”项目“面向资源稀缺语言的实体挖掘及应用研究”（61976016）；“融合谓词框架和语义知识的神经机器翻译研究”(61976015)；“基于语义组合的开放域汉语复述研究”(61876198)。

作者简介 李思琪（1997-），硕士，研究方向为自然语言处理、信息抽取；朱庆（1996-），硕士，研究方向为自然语言处理、信息抽取；陈钰枫（1981-），副教授，研究方向为自然语言处理、机器翻译，E-mail：chenyf@bjtu.edu.cn；徐金安（1970-），教授，研究方向为自然语言处理、机器翻译；张玉洁（1961-），教授，研究方向为自然语言处理、机器翻译。

引用格式 李思琪 , 朱庆 , 陈钰枫 , 等 . 融合多头自注意力的远程监督关系抽取方法 [J]. 情报工程 , 2021, 7(6): 45-57.

doi:10.3772/j.issn.2095-915x.2021.06.004

Abstract: [Objective/ Significance] In the research of relation extraction, distant supervision can construct amount of training corpus in a short time, while introduces a lot of noisy instances because of wrong labels. The existing research work primarily uses attention mechanism to reduce the impact of noise. However, when distributing attention weights, these methods are based on labels containing noise, which would further worsen wrong labelling problem. [Methods/Process] This paper proposes a distant supervision relation extraction method based on multi-head self-attention. Word-level attention weights are obtained by multi-head self-attention, and sentence-level attention weights are obtained based on the semantic representation of head and tail entities, so as to avoid using noisy relation labels as the basis for attention allocation, which reduces the influence of noise. [Results /Conclusions] The experimental results on the public dataset show that the proposed model has a significant improvement compared to all baseline models.
Keywords: Relation extraction; distant supervision; multi-head self-attention; entity feature

引言

关系抽取作为自然语言处理领域中一项重要的基础任务，旨在判断出文本语句中实体对之间存在的语义关系。传统的基于有监督方法的关系抽取依赖于高质量的人工标注数据，需要花费大量的时间和人力。于是，能够在现有的知识库和语料库的基础上，自动标注训练数据的远程监督方法逐渐成为了关系抽取研究任务的热点。

远程监督的概念由 Mintz 等 [1] 于 2009 年提出，其核心思想是：若在现有的知识库中存在某一个实体关系三元组（实体 1，关系 R，实体 2），则认为在语料库中，所有同时包含有实体 1 和实体 2 的文本语句，都表达了关系 R。例如：在知识库中，有一个实体关系三元组（比尔盖茨，创始人，微软），则将所有“比尔盖茨”和“微软”同时出现的句子都标记为“创始人”的关系。通过远程监督思想可以在短时间内获得大量标注好的训练语料，但是由于其假设过于强硬，不可避免地导致了错误标注的噪声问题。例如在句子“在 1992 年，比尔盖茨拥有超过 40% 的微软股票，经过这么多年不断地抛售后，现在他只有 1% 了”中，同时出现了“比尔盖茨”和“微软”两个实体词，但是表达的关系却不是“创始人”的关系。因此，如何降低语料中的噪声数据影响，提升关系抽取的性能成为了当前研究的重点。

1 相关工作

由于远程监督构建数据集的特殊性，导致获得的数据集中存在大量的噪声标签，影响了关系抽取的性能。为了处理噪声问题，有学者提出将远程监督关系抽取看作一个多示例学习（Multi Instance Learning, MIL）任务，将包含有同一实体对的所有句子看作一个实体包，基于实体包来进行实体对的关系抽取 [2]，有效地缓解了错误标注的问题；之后也有学者们提出基于多示例学习的概率图模型 [3-4]。Zeng 等 [5] 提出了一种结合多示例学习思想的分段卷积神经网络（Piece-wise Convolutional Neural Net-

work，PCNNs），基于 At-least-once 的思想，认为在实体包内至少存在一个句子，能够正确的表达实体对之间的关系，选择最优句子作为整个实体包的表示。之后，Lin 等 [6] 认为应该将实体包内所有句子的信息都利用起来，组合作为实体包的表示，并提出使用句子级别的注意力机制来分配实体包内不同句子的权重。Ji 等 [7] 提出使用基于 TransE 思想 [8] 的句子级别注意力机制，将两个实体的差值融合到注意力计算中，并且还加入了知识库中实体的描述信息。Jat 等 [9] 使用了词级别和句子级别两层注意力，进一步降低了远程监督的噪声影响，取得了较好的效果。Yuan 等 [10] 采用非独立非同分布关联嵌入法来捕获袋内句子的相关性，得到更好的包向量表示。Alt 等 [11] 通过选择性注意机制将标准的 Transformer 结构扩展到多实例学习，并在远程监督关系抽取任务上进行微调，减少了显式特征提取以及误差累积的风险。Ye 等 [12] 同时考虑了包内和包间的噪声，分别对句子级和包级噪声进行处理。针对远程监督数据存在的错误标注问题，Huang 等 [13] 基于上下文相关的矫正策略将可能错误的噪声标签修正到正确方向。Shang 等 [14] 利用无监督深度聚类为含噪句子生成可靠标签。Wang 等 [15] 提出了一种无标签方法，即利用类型信息和翻译规律对学习过程进行软监督，不需要额外的降噪模型。另外，有学者还提出应该在关系抽取模型之前就过滤数据集中的噪声，Qin 等 [16] 通过生成对抗训练来去除噪声，处理错误标注的句子。Feng 等 [17] 引入强化学习，将过滤噪声句子建模为一个强化学习决策问题，根据删除一个句子后的关系抽取器性能表现作为强化学习的奖励或者处罚。

当前远程监督关系抽取经过不同研究者的不断努力创新，已经取得了较好的性能表现，但是仍存在三个问题影响了关系抽取的性能提升。（1）远程监督构建的句子中存在大量和关系表示无关的噪声词，根据 Liu 等 [18] 的研究：在 NYT-Freebase 这一远程监督关系抽取经典数据集中，约有 99.4% 的句子都存在无关的噪声词，平均每条句子中都有着 12 个无关词。无关的噪声词影响了关系抽取模型提取到的句子特征的质量，而现有的方法直接以关系标签为依据，计算不同词与关系标签的相关度，进而分配不同词的注意力权重，没有考虑到关系标签会存在大量的噪声，向错误关系标签学习，导致了“越学越错”的问题，影响了关系抽取的性能。（2）现有的方法在生成实体包特征时，直接采用含有噪声的关系标签作为计算注意力权重的依据，忽略了关系标签噪声的负面影响，难以获得合理的实体包特征表示。（3）Ye 等 [19] 的研究表明，句子中的命名实体词往往对于句子的语义表示有较高的影响，而现有的基于词向量和位置向量的句子输入表示忽略了文本句子中的命名实体对于句子表示的重要性，另外头、尾的实体词作为句子中最核心的词，也应受到更多的关注度。

针对以上问题，我们提出了一种新的融合多头自注意力和实体特征的分段卷积神经网络模型（Entity-wised Multi-head Self Attention based PCNNs，简称为 EMSA_PCNN），与 JAT 等人在词级和句子级注意力都依据关系标签分配注意力不同的是，我们在 PCNNs 提取句子特征的基础上，使用多头自注意力来更加合理地分配不同词的贡献度，不需要额外的监督信息，

doi:10.3772/j.issn.2095-915x.2021.06.004

缓解了远程监督的无关噪声词影响句子表示的问题。除此之外，我们在生成实体包时，基于TransE 算法的思想：$e_1 - e_2 \approx r$，认为实体之间的关系可以由头、尾实体进行一定的计算变换得到，采用头、尾实体经过双线性变换后的向量表示作为句子权重的计算依据，通过缩放点积注意力处理，完全不依赖关系标签，进一步缓解了远程监督关系抽取的噪声在生成实体包特征时的负面影响。另外，在输入部分，额外加入了命名实体和核心的头、尾实体词特征，进一步丰富了句子的输入特征表示，有利于关系抽取模型学习到更多有效的语义特征。我们在 NYT-Freebase 数据集上的实验结果表明：相较于基线系统，我们提出的方法性能有了显著提升，验证了提出的方法的有效性。

2 模型架构

如 EMSA_PCNN 模型架构图如图 1 所示，对于实体包 B 中的所有句子，EMSA_PCNN 模型首先将文本句子转换为向量表示（2.1 节融合实体特征的输入层），然后采用卷积层提取文本特征（2.2 节卷积层），再通过多头自注意力机制来合理分配不同词的权重（2.3 节多头自注意力层），然后对文本表示进行分段的最大池化处理获得最终句子表示（2.4 节分段最大池化层），再通过基于头、尾实体双线性变换的注意力计算合理分配不同句子的权重，综合所有句子特征以获得最终实体包 B 的表示（2.5 节融合实体特征的句子特征选择层），最后通过经过 Softmax 激活的全连接层处理，获得实体包 B 对于所有关系类别的得分。

相较于现有的基于 PCNNs 的模型，我们提出的方法在卷积层之后使用多头自注意力机制来分配句子中不同词的权重，并且在生成实体包时，完全不使用含有噪声关系标签的特征信息，使得模型能进一步避免远程监督噪声的影响。

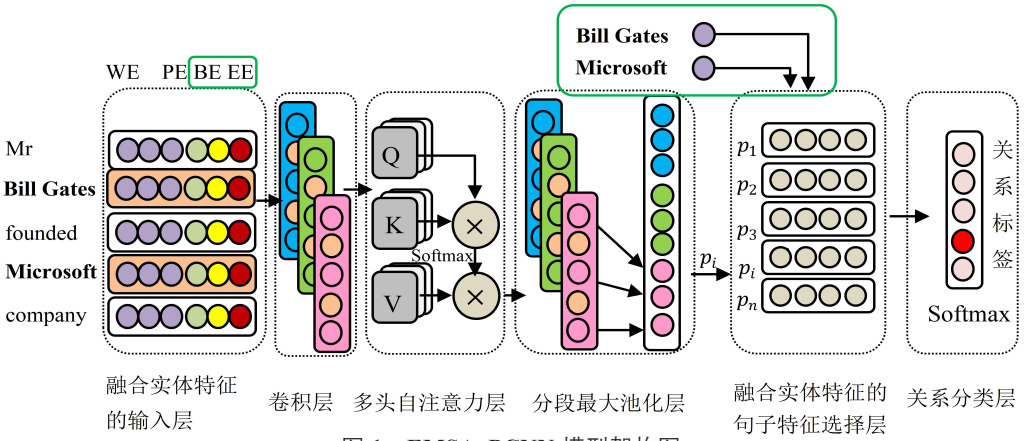

图 1 EMSA_PCNN 模型架构图

2.1 融合实体特征的输入层

2.1.1 词向量

词向量是对词的分布式表示，目的是将文本中的每一个词都转换成低维的数字向量。具体来说，对于包含 n 个词的输入句子 s，可表示为 $s=\{w_1, w_2, w_3, \ldots, w_n\}$，首先将其中每个词

$w_i\{i=1, 2, 3,\ldots, n\}$ 都通过词典映射为数字索引号，其次通过映射矩阵 $\boldsymbol{WE}\in\mathbb{R}^{|V|\times d_w}$ 将每个索引号都转换成向量形式 $v_i(i=1, 2, 3,\ldots, n)$，词向量的维度为 d_w，$|V|$ 是词典大小。

2.1.2 位置向量

句子中离实体词越近的词对于句子表示的贡献越大，通过每个词与两个实体词的相对距离大小可以获得位置向量，弥补了 CNN 对于位置信息编码能力较弱的缺点。具体来说，句子中第 i 个词与两个实体的相对位置大小分别用 $P1_i(i=1, 2, 3,\ldots, n)$ 和 $P2_i(i=1, 2, 3,\ldots, n)$ 表示，再通过两个向量矩阵 $\boldsymbol{PE}_1$ 和 $\boldsymbol{PE}_2$，将数值映射为低维向量 $\boldsymbol{p1}_i(i=1, 2, 3,\ldots, n)$ 和 $\boldsymbol{p2}_i(i=1, 2, 3,\ldots, n)$，位置向量的映射维度为 d_p。相对位置如图 2 所示：“technology”这个词到头实体词“Bill Gates”和尾实体词“Microsoft”的相对距离分别是“4”和“-6”。

图 2　相对距离示例

2.1.3 命名实体向量

文本中存在不同类型命名实体词，将命名实体的开头词标记为“命名实体类型 -B”，命名实体的后续词标记为“命名实体类型 -I”，不是命名实体的词的标记为“O”。如图 3 所示的一个句子：“On July 28, Bill Gates sold 2 million Microsoft shares”，其中的“July”和“28”分别被标记为 DATE-B 和 DATE-I 类型，“Bill”和“Gates”分别被标记为 PER-B 和 PER-I 类型，“Microsoft”被标记为 ORG-B 类型。给定输入句子 $s=\{w_1, w_2, w_3, \ldots, w_n\}$，将其命名实体标签（BIO 类型）$NE=\{ne_1, ne_2, ne_3, \ldots, ne_n\}$ 转换成向量表示 $\boldsymbol{BE}=\{\boldsymbol{bio}_1, \boldsymbol{bio}_2, \boldsymbol{bio}_3, \ldots, \boldsymbol{bio}_n\}$，命名实体向量的维度为 d_{bio}。

词	On	July	28	,	Bill	Gates	sold	2	millon	Micrsoft	shares
标签	O	DATE-B	DATE-I	O	PER-B	PER-I	O	O	O	ORG-B	O

图 3　命名实体类型示例

2.1.4 头、尾实体向量

头、尾实体词是文本中最核心的两个词，在输入阶段强化两者的影响，将头、尾实体词向量的差值作为其特殊特征，即 $\boldsymbol{ee}=\boldsymbol{v}_{e_1}-\boldsymbol{v}_{e_2}$。

对于输入句子 $s=\{w_1, w_2, w_3, \ldots, w_n\}$，通过上述的词向量、位置向量、命名实体向量和头、尾实体向量，将 s 转换为一个二维矩阵向量 $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_n\}$，$\boldsymbol{X}\in\mathbb{R}^{n\times d_x}$，其中 $\boldsymbol{x}_i=\{\boldsymbol{v}_i; \boldsymbol{p1}_i; \boldsymbol{p2}_i; \boldsymbol{bio}_i; \boldsymbol{ee}\}$，$n$ 是输入文本句子长度，d_x 是模型的输入向量维度，表示为 $d_x=2\times d_w+2\times d_p+d_{bio}$。$\boldsymbol{X}$ 将作为后续卷积层的输入。

2.2 卷积层

卷积神经网络（Convolutional Neural Network，CNN）是 NLP 领域提取文本特征的常用模型之一，能够获取文本中不同的 n-gram 特征，并且计算时可以并行处理。

给定输入句子矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_n\}$，其中 $\boldsymbol{x}_i\in\mathbb{R}^{d_x}$，$\boldsymbol{x}_{i:j}$ 表示 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的拼接矩阵，权重矩阵 $\boldsymbol{W}_c\in\mathbb{R}^{w\times d_x}$ 卷积操作的卷积核，首先通过大小为 w 滑动窗口与 $\boldsymbol{X}$ 进行卷积运算，超出句子最大长度的部分用 0 填充。通过卷积处理后，可以得到文本特征 $\boldsymbol{c}=\{c_1, c_2, c_3, \ldots, c_{n-w+1}\}$，$c_j$ 的计

doi:10.3772/j.issn.2095-915x.2021.06.004

算如公式 (1) 所示：

$$c_i=\boldsymbol{W}_c\boldsymbol{x}_{i:i+w-1}+b_c,\ 1\le i\le n-w+1 \quad (1)$$

采用 Swish 函数作为激活函数，其计算如公式 (2)、公式 (3) 所示，其中 β 为可学习的超参数。

$$\text{Swish}(x)=x\cdot\text{Sigmoid}(x\beta) \quad (2)$$

$$\text{Sigmoid}(x)=1/(1+e^{-x}) \quad (3)$$

使用多个卷积核进行卷积运算以获得文本的多粒度特征信息，卷积核集合为 $\boldsymbol{W}_c=\{\boldsymbol{w}_c^1, \boldsymbol{w}_c^2, \ldots, \boldsymbol{w}_n^k\}$，进行 k 次卷积处理。获得文本特征表示 $\boldsymbol{C}=\{\boldsymbol{c}^1, \boldsymbol{c}^2, \ldots, \boldsymbol{c}^k\}$，$\boldsymbol{C}\in\mathbb{R}^{k\times(n+w-1)}$。

2.3 多头自注意力层

远程监督构建的训练句子中往往存在大量和关系表示无关的噪声词，现有的方法在处理这些噪声词时，往往直接采用远程监督的关系标签作为计算注意力的依据，重新分配词的注意力权重，没有考虑到远程监督关系标签的噪声问题对于词权重分配的影响。于是，我们选择使用多头自注意力机制来处理这一问题，不需要关系标签作为注意力分配依据，通过计算词与词之间的相关性，动态地分配不同词的权重，过滤无关词的负面影响。

多头自注意力（**M**ulti-head **S**elf **A**ttention，MSA）机制来源于 Google 所提出的 Transformer 模型 [20]，在机器翻译任务中被广泛使用，性能表现十分突出。MSA 是一种特殊的内部注意力计算方法，对一个序列本身计算并且重新分配各个位置合适的注意力权重，以获得更合理的特征表示。MSA 的计算过程主要包括自注意力计算和多头拼接两个部分组成。

2.3.1 自注意力计算

自注意力是指将句子中每个词都与其他词计算注意力权重，单次自注意力计算过程为：将卷积层获得的特征 $\boldsymbol{C}$ 通过三个不同的线性变换处理，分别得到 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 三个矩阵，再通过缩放点积注意力计算，得到单次自注意力后的句子表示，缩放点积注意力计算过程表示为公式 (4)，其中 d_k 表示矩阵 $\boldsymbol{K}$ 的维度。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (4)$$

2.3.2 多头拼接

多头拼接部分指进行多次自注意力计算，将多次计算结果进行拼接，得到最终的句子表示，以综合在多个向量子空间内的权重表示。

$$\boldsymbol{head}_i=\text{Attention}(\boldsymbol{C}\boldsymbol{W}_i^q,\boldsymbol{C}\boldsymbol{W}_i^k,\boldsymbol{C}\boldsymbol{W}_i^v),1\le i\le h \quad (5)$$

$$\boldsymbol{H}=\text{Concat}\{\boldsymbol{head}_1, \boldsymbol{head}_2, \ldots, \boldsymbol{head}_h\}\boldsymbol{W}_{sa} \quad (6)$$

其中 $\boldsymbol{W}_i^q,\boldsymbol{W}_i^k,\boldsymbol{W}_i^v,\boldsymbol{W}_{sa}\in\mathbb{R}^{k\times k}$，经过多次自注意力运算并拼接后，可以得到新的特征 $\boldsymbol{H}=\{\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \ldots, \boldsymbol{h}_k\}$。

2.4 分段最大池化层

在 NLP 领域，往往使用最大池化来提取句子表示中最显著的特征，并且降低参数量，但是在关系抽取领域，最大池化选择粒度过大，难以获取到实体对周围的重要信息。由 Zeng 等 [5] 提出了分段最大池化，根据句子中两个实体的位置将句子分为三个部分，对三个部分分别进行最大池化，对于多头自注意力层得到的特征表示 $\boldsymbol{H}=\{\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \ldots, \boldsymbol{h}_k\}$，每个 $\boldsymbol{h}_i$ 会被分为 $\{\boldsymbol{h}_{i1}, \boldsymbol{h}_{i2}, \boldsymbol{h}_{i3}\}$，则分段最大池化定义为：

$$p_{ij}=\max(\boldsymbol{h}_{ij}),\ 1\le i\le k,\ 1\le j\le 3 \quad (7)$$

将所有的 $\boldsymbol{p}_i=\{p_{i1}, p_{i2}, p_{i3}\}$ 进行拼接，再采用 Swish 函数激活，可得到句子的最终表示 $\boldsymbol{P}\in\mathbb{R}^{3k}$。

2.5 融合实体特征的句子特征选择层

现有的研究在选择句子特征进而生成实体包特征时，采用的句子级别注意力机制是以关系标签为分配权重依据的，这种方法已取得了良好的效果，但是关系标签中所存在的噪声问题仍然对实体包级别的特征有负面影响。于是，我们基于 TransE 算法的思想，认为关系能够通过头、尾实体之间的计算得到，采用头、尾实体的词向量（$\boldsymbol{v}_{e_1}$,$\boldsymbol{v}_{e_2}$）经过双线性变换所学习得到的向量表示 $\boldsymbol{r}$ 作为注意力计算的依据。

$$\boldsymbol{r}=\boldsymbol{v}_{e_1}{}^{T}\boldsymbol{W}_e\boldsymbol{v}_{e_2}+b_e \tag{8}$$

对于实体包 B 中第 j 个句子特征 $\boldsymbol{P}_j$ 的权重 $\boldsymbol{a}_j$ 计算如下所示：

$$a_j=\mathrm{softmax}\left(\frac{\boldsymbol{P}_j^T\boldsymbol{r}}{\sqrt{d_k}}\right) \tag{9}$$

实体包 B 最终的特征表示为 $\boldsymbol{E}=\Sigma_i a_i\boldsymbol{P}_i$。

再将实体包 B 的表示 $\boldsymbol{E}$ 通过 softmax 分类器，预测最终的关系标签，使用交叉熵函数计算损失：

$$J(\theta)=-\frac{1}{r}\sum_{i=1}^{r}t_i\log(y_i)+\lambda\|\theta\|_F^2 \tag{10}$$

3 实验结果

3.1 实验数据及评测标准

我们采用 NYT-Freebase 作为实验数据集，该数据集由 Riedel 等[2]于 2010 年构建并且开源，是采用远程监督方法将 Freebase 知识库和纽约时报（New York Times，NYT）语料库对齐得到的，包含 NA 在内的共 53 种关系类别，NA 即表示实体对间没有关系。将 NYT 语料中的 2005—2006 年的文本作为训练集，2007 年的文本作为测试集。其中，测试集的实体关系标签是人工标注的，可靠性较高，共 172 448 条句子，96 678 个实体对以及 1950 个关系事实；训练集则完全由实体关系三元组匹配得到，存在噪声，共 522 611 条句子，281 270 个实体对以及 18 252 个关系事实。

同现有的研究一样，我们采用 held-out evaluation 方法来评估我们提出的关系抽取模型。通过模型的 PR 曲线（Precision Recall Curve）、P@N（Precision@Top N）以及 AUC 值（Area Under Curve）的表现来评估模型的性能。

3.2 实验设置

我们提出模型使用的词向量是通过 Word-2Vec 在 NYT 语料上预训练得到的，而位置向量和命名实体向量则是随机初始化生成的。我们采用网格搜索法来确定模型最优的超参数设置。其中卷积核数目取值 $k\in\{100, 200, 230, 256, 300\}$，滑动窗口大小取值 $w\in\{3, 4, 5\}$，批处理大小取值 batch $\in\{64, 128, 160\}$，神经元随机失活率取值 dropout $\in\{0.3, 0.5, 0.7\}$，学习率取值 lr $\in\{0.001, 0.01, 0.05, 0.1\}$，多头自注意力的头数取值 h $\in\{1, 3, 5, 7\}$。最终的超参数设置为：卷积核数目为 230，滑动窗口大小为 3，批处理为 160，神经元随机失活率为 0.5，学习率为 0.01，多头自注意力头数为 5，采用的优化算法为 Adadelta。

3.3 实验结果及对比分析

我们将提出的 EMSA_PCNN 模型与现有的多个模型进行对比：

doi:10.3772/j.issn.2095-915x.2021.06.004

（1）Mintz[1]：由 Mintz 首次提出远程监督的概念，通过多分类 logistic 回归分类器处理关系抽取。

（2）MultiR[3]：由 Hoffman 提出的采用多示例学习的概率图模型。

（3）MIML[4]：由 Surdeanu 提出的处理多示例多标签问题的概率图模型。

（4）PCNNs+MIL[5]：由 Zeng 提出的结合多示例学习和 PCNN 模型，选择实体包内最优句子来代表整个实体包。

（5）PCNNs+ATT[6]：基于 PCNN 模型，由 Lin 提出的采用句子级别注意力机制的方法，根据关系标签分配实体包内不同句子的权重。

（6）APCNNS[7]：基于 PCNN 模型，由 Ji 提出的使用两实体之差辅助的信息计算句子级别注意力机制的方法。

（7）BGWA[9]：由 Jat 提出，采用 BiGRU 提取特征，基于关系标签分别使用词级别和句子级别两种注意力的方法。

根据图 4 和表 1 可以看出：

（1）相较于基于机器学习的传统方法 Mintz、MIML、MultiR，其他的基于深度学习的方法效果明显更好，在 PR 曲线、P@N 和 AUC 值等表现都更好，这说明传统的人工设计的特征在性能上难以与深度学习自动提取到的特征相比较。

（2）EMSA_PCNN 在 PR 曲线和各项指标上都优于 PCNNs+MIL、PCNNs+ATT 以及 APCNNS 模型，AUC 值比 APCNNS 模型提升了 2.5 个百分点，P@Mean 值提升了 6.2 个百分点。这是由于 EMSA_PCNN 不光在句子级别进行了降噪处理，还降低了词级别的无关噪声，进一步缓解了远程监督的错误标注的问题。

（3）EMSA_PCNN 与 BGWA 相比，PR 曲线与各项指标都更优，在 AUC 值上提升了 2.2 个百分点，在 P@Mean 值上提升了 5.9 个百分点。这是因为 EMSA_PCNN 在词级别降噪时，使用了多头自注意力来分配权重，避免了噪声关系标签的负面影响，另外在句子级别降噪时，EMSA_PCNN 使用了头、尾两实体经过双线性变换的向量表示作为计算注意力的依据，进一步降低了远程监督错误标注的影响。

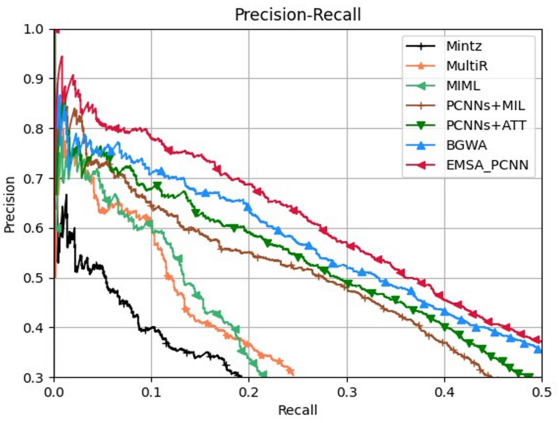

图 4　**EMSA_PCNN** 与基准模型的 **PR** 曲线图

表 1　不同模型的 **P@N**、**AUC** 值对比（**%**）

模型	P@100	P@200	P@300	P@Mean	AUC
Mintz[1]	51.8	50.0	44.8	48.9	10.6
MIML[3]	70.9	62.8	60.9	64.9	12.0
MultiR[4]	70.2	65.1	61.7	65.7	12.6
PCNNs+MIL[5]	72.3	69.7	64.1	68.7	32.5
PCNNs+ATT[6]	76.2	73.1	67.4	72.2	34.1
APCNNS[7]	76.3	74.2	69.4	73.3	35.7
BGWA[9]	75.2	74.1	71.4	73.6	36.0
EMSA_PCNN	**83.2**	**79.6**	**75.7**	**79.5**	**38.2**

此外，本文还与 2019 年最先进的两种关系抽取模型进行了性能对比，这两种方法使用的是 Liu 等 [18] 提供的数据集，本文称为

NYT-SoftRE，与先前介绍的 NYT-Freebase 不同的是 NYT-SoftRE 是在 NYT-Freebase 基础上增加了约 5 万条训练句子，而测试句子保持不变。选取的关系抽取模型有：(1)PCNN+BAG-ATT：Ye 等 [12] 在句子级别注意力的基础上，进一步增加实体包级别的注意力机制，降低了实体包级别的噪声；(2)PCNN+WN：Yuan 等 [10] 提出在向量表示阶段加入注意力计算，在句子选择部分认为实体包内的句子不是独立同分布的，将最优句和其他句子计算相似度得到注意力权重的方法。为了更加直观的看出不同方法在 NYT-SoftRE 上的性能对比，除了这两种方法之外还加入了 PCNNs+ATT 作为对比模型，本文方法和上述几种方法的 AUC 值对比如表 2 所示：

表 2 基于 NYT-SoftRE 数据集的模型的 AUC 值对比 (%)

模型	AUC
PCNNs+ATT[6]	38.9
PCNNs+BAG-ATT[12]	42.2
PCNN+WN[10]	**43.0**
EMSA_PCNN	42.9

根据表 2 可知，在 NYT-SoftRE 数据集上，相较于 PCNNs+ATT 模型，本文提出的 EMSA_PCNN 模型在 AUC 值表现上提升了 4 个百分点，验证了本文方法的有效性。相较于加入实体包级别注意力的 PCNN+BAG-ATT 模型，EMSA_PCNN 在性能上有 0.7 个百分点的提升。而相较于 PCNN+WN 模型，本文提出的 EMSA_PCNN 模型的 AUC 值表现仅低 0.1 个百分点，说明我们提出的方法也具有较好的竞争力。在以后的研究工作中，我们可以进一步探索实体包中不同句子之间的联系，而不是将所有的句子看作独立的个体，从而提升关系抽取的性能。

4 消融实验

4.1 融合多头自注意力的性能评测

为了验证多头自注意力机制对于远程监督降噪的有效性，我们设计了采用 PCNN 为句子编码器，使用一般的句子级别注意力，多种方式计算词的注意力，进行对比实验。(1)PCNNs+ATT：未采用任何词级别降噪的方法。(2)WA+PCNN+ATT：直接采用关系标签作为监督信息计算词级别注意力的方法。(3)MSA_PCNN：不采用任何监督信息，使用多头自注意力计算词级别注意力的方法。

从图 5 和表 3 可以看出，加入普通词级别注意力的 WA+PCNN+ATT 模型在性能上有一部分提升，说明无关噪声词的确影响了关系抽取的性能。另外，MSA_PCNN 模型的表现更好，在 PR 曲线、P@N 以及 AUC 值上都较于 WA+PCNN+ATT 有显著提升，这是因为 MSA_

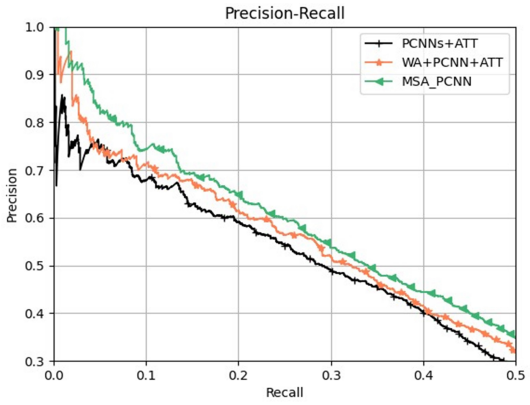

图 5 不同词级别注意力机制的性能对比

doi:10.3772/j.issn.2095-915x.2021.06.004

表 3　不同词级别注意力机制的 P@N，AUC 值对比（%）

模型	P@100	P@200	P@300	P@Mean	AUC
PCNNs+ATT	76.2	73.1	67.4	72.2	34.1
WA+PCNN+ATT	76.2	72.1	70.4	72.9	35.2
MSA_PCNN	**84.2**	**79.1**	**74.4**	**79.2**	**37.5**

PCNN 采用的多头自注意力避免了噪声关系标签的影响，另外“多头”机制能够综合多个向量子空间的权重信息，进一步降低了词级别的噪声，验证了多头自注意力的有效性。

4.2　融合实体特征的性能评测

为了验证融合实体特征的有效性，我们设计了同样采用 PCNN 为句子编码器，不同方法计算句子级别注意力的方法。(1)PCNNs+ATT：以关系标签为依据，计算句子级别注意力的方法；(2)APCNNS：以两实体之差作为辅助关系标签的信息，计算句子级别注意力的方法；(3)PCNNs+EATT：我们提出的使用头、尾实体经过双线性变换的向量表示作为依据，采用缩放点积注意力分配句子权重的方法。

根据图 6 和表 4 可知，采用头、尾实体经过双线性变换的 PCNNs+EATT 模型相较于 PCNNs+ATT 和 APCNNS 模型来说，在 PR 曲线、P@N 和 AUC 值上表现都有较好提升，验证了融合实体特征在句子级别注意力计算的有效性。

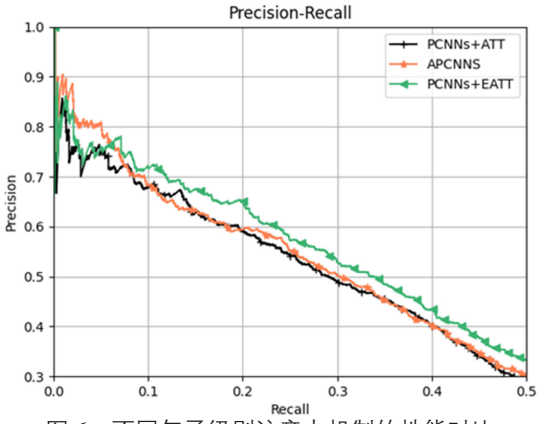

图 6　不同句子级别注意力机制的性能对比

表 4　不同句子级别注意力机制的 P@N，AUC 值对比（%）

模型	P@100	P@200	P@300	P@Mean	AUC
PCNNs+ATT	76.2	73.1	67.4	72.2	34.1
APCNNS	**76.3**	**74.2**	69.4	73.3	35.7
PCNN+EATT	**76.3**	74.1	**71.4**	**73.9**	**36.1**

另外，为了验证在输入部分融合实体特征信息的有效性，我们同样也以 PCNN 为编码器，在句子选择时采用多示例学习（MIL）和普通句子级别注意力（ATT），根据输入部分是否加入实体特征（E），分别记为：(1)E+PCNNs+MIL 和 PCNNs+MIL,(2)E+PCNNs+ATT 和 PCNNs+ATT。

根据图 7 可以看出，加入命名实体特征和实体词特征的模型在 PR 曲线表现中都优于未加入的模型，验证了在输入层加入额外特殊词特征的有效性。

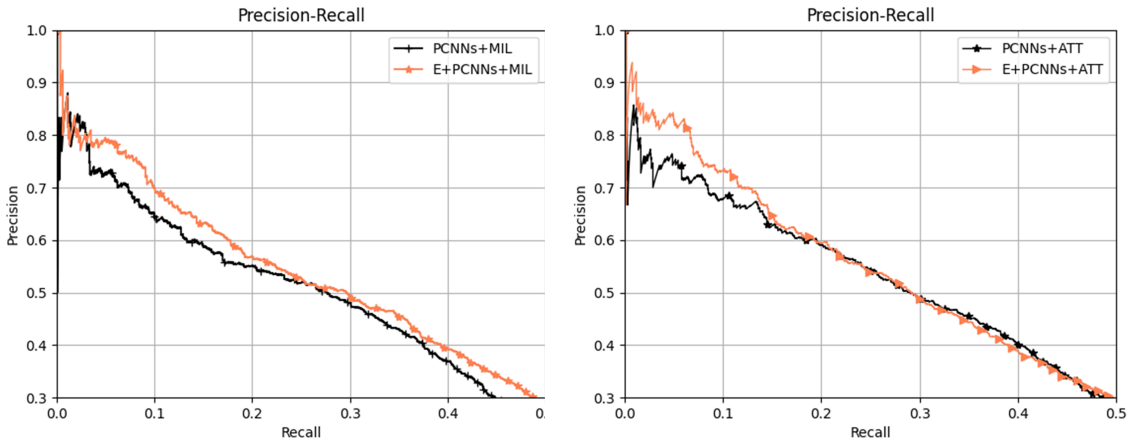

图 7　输入层融合特殊词特征的性能对比

5　实例分析

为了更直观地解释多头自注意力机制能够缓解词级别噪声的原因，我们从测试集中选取了一个实例，分别对 5 个头的注意力权重进行了可视化的展示，如图 8 所示。由于关系抽取任务旨在抽取两个实体间的语义关系，因此我们仅关注实体对（Boston，Back Bay）与句内其他词之间相关性，其中颜色越深表示所分配的权重越大，即相关性越高。

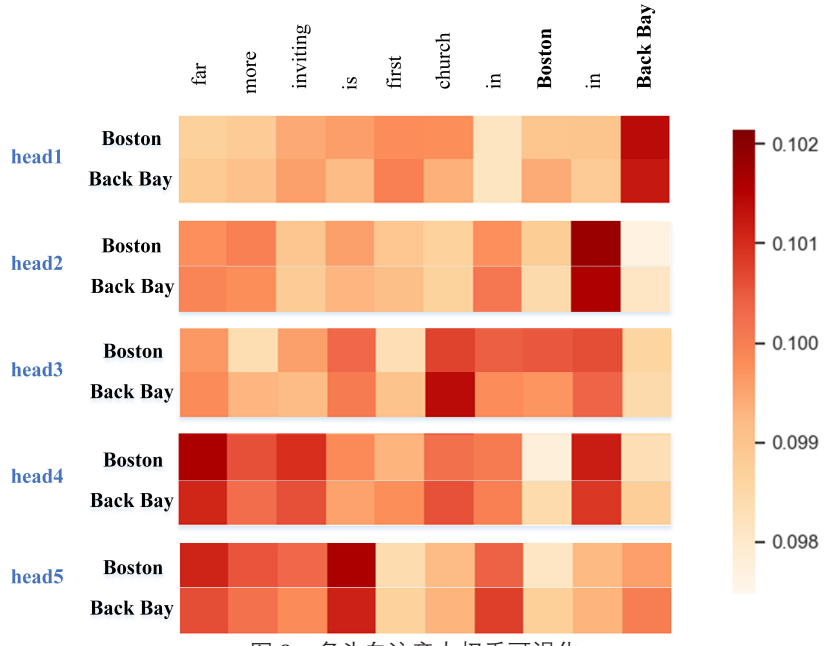

图 8　多头自注意力权重可视化

通过权重分布可以看出，多头自注意力机制可以从5个不同的表示子空间学习相关信息，为“more”“inviting”“first”等无关词分配低权重，同时也为“in”“church”等相关单词分配了高权重，为预测“Boston”与“Back Bay”之间的 /**contains**/ 关系提供了有效信息。

doi:10.3772/j.issn.2095-915x.2021.06.004

由此可见，多头自注意力机制可以降低无关词的噪声信息的影响，从词级别缓解噪声传播，从而提升关系抽取性能。

6 结语

针对当前词级别、句子级别降噪的注意力计算时都依据含有噪声的关系标签的问题，本文提出了一种融合多头自注意力和实体特征的分段卷积神经网络方法，并将其应用到远程监督关系抽取任务中。在词级别降噪时，采用多头自注意力来更加合理地分配每个词的贡献度，降低句子中无关的噪声词对于句子表示的负面影响；在句子级别降噪时，采用头、尾实体经过双线性变换的隐层向量作为注意力计算的依据，两者注意力计算时都没有以关系标签作为依据，降低了关系标签中噪声对于注意力权重分配的不利影响，进一步缓解了远程监督的错误标注的问题。实验结果表明：我们提出的模型优于所有的基线模型，达到了最好的效果。在以后的研究中，我们将进一步融合知识库的其他有用信息，以及预训练语言模型的知识来探索进一步提升关系抽取性能的方法。

参 考 文 献

[1] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009:1003-1011.

[2] Riedel S, Yao L, Mccallum A. Modeling relations and their mentions without labeled text[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg. 2010:148-163.

[3] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. 2011:541-550.

[4] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction[C]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics. 2012:455-465.

[5] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1753-1762.

[6] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:2124-2133.

[7] Ji G, Liu K, He S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017:3060-3066.

[8] Bordes A, Usunier N, Garcia-duran A, et al. Translating embeddings for modeling multi-relational data[C]. Advances in Neural Information Processing Systems. 2013:2787-2795.

[9] Jat S, Khandelwal S, Talukdar P. Improving distantly supervised relation extraction using word and entity

based attention[J]. arXiv preprint arXiv:1804.06987, 2018.

[10] Yuan C, Huang H, Feng C, et al. Distant Supervision for Relation Extraction with Linear Attenuation Simulation and Non-IID Relevance Embedding[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2019:7418-7425.

[11] Alt C, Hubner M, Hennig L. Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:1388-1398.

[12] Ye Z X, Ling Z H. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:2810-2819.

[13] Huang X, Zhang B, Ye Y, et al. A Noise Adaptive Model for Distantly Supervised Relation Extraction [C]. Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing. 2020:519-530.

[14] Shang Y, Huang H, Mao X, et al. Are Noisy Sentences Useless for Distant Supervised Relation Extraction? [C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 8799-8806.

[15] Wang G, Zhang W, Wang R, et al. Label-free distant supervision for relation extraction via knowledge graph embedding [C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2018:2246-2255.

[16] Qin P, Xu W, Wang W Y. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction [C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:496-505.

[17] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data [C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018:5779-5786.

[18] Liu T, Wang K, Chang B, et al. A soft-label method for noise-tolerant distantly supervised relation extraction [C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017:1790-1795.

[19] Ye W, Li B, Xie R, et al. Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data [C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1351-1360.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. Advances in Neural Information Processing Systems. 2017: 5998-6008.