基于位置敏感 Embedding 的中文命名实体识别

鲁亚楠，孙锐，姬东鸿

（武汉大学计算机学院，湖北 武汉 430072）

摘 要: 在基于条件随机场和表示学习算法的中文命名实体别任务中，为了缓解现有 Embedding 模型自动学习的特征所存在的语义表示偏差给命名实体识别带来的问题，本文一个基于位置敏感 Embedding 的中文命名实体识别方法。该方法将位置信息融入 Embedding 模型中，采用多尺度聚类方法抽取不同粒度的 Embedding 特征，通过条件随机场来识别中文命名实体。实验结果表明，与传统方法进行比较，F 值提高了 2.85%。

关键词: 命名实体识别；表示学习；Embedding；多尺度聚类；条件随机场

中图分类号： TP391.1 **文献标志码**：A **文章编号**：

Chinese Named Entity Recognition based on Position-sensitive Embedding

Lu Yanan, Sun Rui, Ji Donghong

(Computer School, Wuhan University, Wuhan Hubei 430072, China)

Abstract: In the task of Chinese named entity recognition based on conditional random fields and representation learning, in order to reduce semantic bias of feature learned automatically by embedding model, this paper presented a Chinese named entity recognition method based on position-sensitive embedding model. This method applied the position information to the embedding model and used multi-scale word clustering to extract different size features. And then recognize Chinese named entity with conditional random fields. The experiment shows that, this method improved the F-score by 2.85, compared to traditional methods.

Key words: named entity recognition; representation learning; embedding; multi-scale clustering; conditional random fields

0 引言

中文命名实体识别任务的目标是识别文本中的人名、地名、机构名等实体名称。它是进行机器翻译、知识图谱构建、信息抽取、自动摘要、语义分析、自动问答等高层任务的基础，在中文信息处理中占据重要地位。但是开放的实体数量庞大、不可枚举、歧义性大，并且没有严格统一的规范。这些问题给命名实体识别带来了巨大挑战。

目前，命名实体识别以统计方法为主，把命名实体识别任务看作序列标注问题。由于条件随机场（CRFs）综合了隐马尔科夫模型（HMM）和最大熵模型（MaxEnt）的优点，在序列标注方面性能优于其他模型，因此 CRFs 在命名实体识别中研究中得到广泛使用[1]。

但是基于条件随机场的命名实体识别性能好坏十分依赖人工特征，针对不同领域的命名实体需要抽取不同的特征，这需要大量的语言学知识。为了尽量减少语言学知识的依赖和避免繁琐的特征工程，研究人员开始将深度学习应用到命名实体识别任务，郭江将 Embedding 模型[2]学习到的特征用在条件随机场模型中提高命名实体识别的性能[3,4]，取得了较好的性能。由于 Embedding 模型[2]利用上下文来表示目标单词的语义，目标单词最相似的词通常是它的上下文相关词，因此会导致语义表示偏差。这种语义表示偏差，在命名实体识别时会产生负面影响。例如“武大”和“樱花”由于经常出现在一起，采用 Embedding 表示时产生较高的语义相似度。如果“武大”识别为机构名，“樱花”被识别为机构名的概率很大。导致这一问题的主要原因是由于在 Embedding 学习时只考虑目标单词的上下文词，并没有考虑其上下文的位置关系。

基于此，本文提出一种位置敏感的 Embedding 的命名实体模型，在 Embedding 学习时引入位置信息，利用位置信息来纠正语义表示偏差，提高 Embedding 语义表示质量。通过把 Embedding 多尺度离散化，充分利用 Embedding 模型学习到的特征，使用条件随机场模型识别中文命名实体。本文提出的中文命名实体识别方法比基线 F 值高 2.85%。

1 相关工作

命名实体识别最早出现在 MUC-6 会议上，主要是识别人名、地名、机构名，当时几乎所有的系统都是采取规则方式的[5]，国内学者王宁根据公司的结构特征总结出规则库对金融领域的公司名进行识别[6]，罗智勇提出一种基于可信度的人名识别[7]。这些基于规则的方法能够在小数据集上取得不错的结果，识别准确率高。但此类方法存在很大的局限性，不同的领域需要总结不同的规则，十分依赖专家总结的规则，不具有通用性，在大数据集上表现不好，召回率低等。

随着大规模语料库的增加，基于统计的方法逐渐成为研究热点。文献[8]提出一种基于隐马尔科夫模型的命名实体识别。隐马尔科夫模型是生成式模型，假设中心词只与前 n 个词有关，与之后的词无关，导致无法充分的利用上下文信息。文献[9]利用最大熵模型融合多种特征识别命名实体，该模型存在标注偏置问题。条件随机场模型的提出解决了标注偏置问题和长距离关联特征问题[1]，成为主流的命名实体识别方法[10]。

基于统计的方法非常依赖语言学知识和繁琐的特征工程[1,9,11,12]。深度学习得益于自动地学习对目标任务有效的特征，可以不依赖特征工程和领域知识达到现有系统的性能[13]，因此利用深度学习来提高自然语言处理性能是最近

收稿日期：修回日期： **基金项目：**国家自然科学基金重点项目(61133012)；国家自然科学基金面上(61373108)

作者简介: 鲁亚楠(1989-)，男，河南开封人，硕士，主研方向:自然语言处理，深度学习，E-mail：luyanan@whu.edu.cn；孙锐，博士，主研方向：自然语言处理；姬东鸿，教授，博士生导师；

几年的研究热点，其中 Embedding 模型是深度学习在自然语言领域发挥作用的基础。文献[3,4]中把 Embedding 学习到的特征应用在 CRFs 中来提高自然语言处理任务的性能。

2 模型算法

本文将中文命名实体识别作为序列标注问题来解决，首先利用大规模语料学习通过特征学习算法自动学习词语的语义特征，本文基于位置敏感 Embedding 模型进行语义特征学习，通过多尺度词聚类算法，获取词语的多尺度聚类特征。然后将自动学习到的多尺度词聚类特征应用在条件随机场模型上进行中文命名实体识别。整体系统框架如图 1 所示。

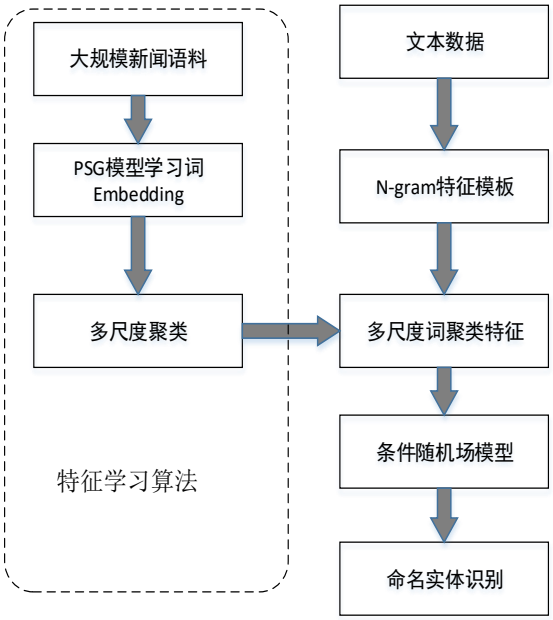

图1 中文命名实体识别框架

2.1 特征学习-Embedding模型

传统统计机器学习模型解决自然语言处理问题时，需要首先通过手工方法抽取特征，通常是要将词进行 one-hot 表示，即每个词表示为一个词表大小的特征向量，向量只有一位是 1，其余都是 0。这种表示会带来两大问题，特征向量极度高维稀疏；无法表示词之间的关系。这导致机器学习模型容易过拟合，并且泛化能力低。

深度学习应用在自然语言处理时，通过 Embedding 模型自动化学习词的语义特征来解决上述问题。Embedding 模型是一种将词进行语义分布式表示的方法，即每个词表示是一个低维稠密向量，语义越相似的词，其表示在 Embedding 空间中距离越接近。该模型利用分布式表示假设，相似的词具有相似的上下文，通过上下文预测窗口中心词或者通过窗口中心词预测上下文来训练模型。文献[1,14]提出了一种快速有效训练 Embedding 模型的方法，用上下文预测窗口中心词（CBOW 模型）或者利用窗口中心词预测上下文（Skip-gram 模型）。由于两种模型基于同一个分布式假设，并且使用相似的网络结构，通过实验验证两种模型训练的 Embedding 质量相差不大。为方便起见，本文基于 Skip-gram 模型进行研究。

2.1.1 Skip-gram 模型

Skip-gram 模型(SG 模型)的学习目标是通过窗口中心词预测上下文的概率。假设窗口大小为 c，窗口中心词为 w_t，上下文 $u[t-c,t-c+1,\ldots,t+c-1,t+c]$。为了简化计算，Mikolov 假设上下文的每个词之间相互独立，给定词 w_t，上下文的概率 $p(u\mid w_t)$ 通过公式(1)求得。通过对语料的所有窗口进行极大似然估计(公式 2)，利用自适应随机梯度下降方法来学习模型参数。由于输出层的大小是整个词表的大小，为了降低模型的时间复杂度，该模型使用文献[14]的负采样算法来训练模型。

$$p(u\mid w_t)=\prod_{j\in[-c,0)\cup[0,c]} p(u_{i+j}\mid w_t) \tag{1}$$

$$\ell=\prod_{w_t\in C}\prod_{u\in \mathrm{Context}(w)} p(u\mid w_t) \tag{2}$$

2.1.2 位置敏感 Skip-gram 模型

由于没有考虑上下文词的位置关系 Skip-gram 模型的会导致语义表示偏差，即上下文相关词，语义相似度很高，为了解决这种问题，本文在 Skip-gram 的基础上融入位置信息，提出位置敏感的 Skip-gram 模型(PSG 模型)。该模型基于假设 1。

假设 1：窗口中心词为 w_t，上下文词 u_{t+j} 相对 w_t 的距离为 j，w_t 对 u_{t+j} 的预测权重为 w_{t+j}，那么在词 w_t 条件下出现 u_{t+j} 的概率通过公式 4 计算。

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{3}$$

$$p(u_{t+j}\mid w_t)=\sigma((w_{t+j}\otimes\vec{u_t})\cdot\vec{w_t}) \tag{4}$$

$$\sigma((w_{t+j}\otimes\vec{u_t})\cdot\vec{w_t})=\sigma(\overrightarrow{pu_{t+j,j}}\cdot\vec{w_t}) \tag{5}$$

公式四中 $\otimes$ 表示向量逐位相乘，通过对公式 4 进一步简化，将 $w_{t+j}\times\vec{u_t}$ 合并为 $\overrightarrow{pu_{t+j,j}}$，$\overrightarrow{pu_{t+j,j}}$ 表示上下文中词 u_{t+j} 在位置 j 的 Embedding。同一个词在不同的位置由不同的 Embedding 表示，因此每一个词由一个位置 Embedding 矩阵表示。

图 2 为 PSG 模型中窗口为 2 的词位置 Embedding 矩阵[1]，词“发”的 Embedding 是一个由四个向量组成的矩阵，在“发明”中，窗口中心词是“明”，是用-1 位置的词向量，在“发工资”中，窗口中心词为“资”，使用-2 位置的词向量。

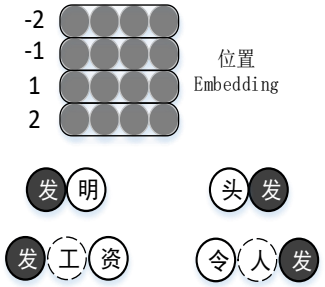

图2 词位置Embedding矩阵

PSG 模型如图 3 所示，该模型窗口周围的词的 Embedding 由一个矩阵表示，图中空心圆的向量表示当前位置词的 Embedding。$pc_{t-2,-2}$ 表示文本中位置为 t-2 的上下文词，在相对窗口中心-2 位置时的 Embedding。

PSG 模型中，窗口中心预测上下文的概率计算方法用公式(6)表示，其中 t 表示句子中当前位置，w_t 表示当前位置的词，j 表示距离窗口中心词 w_t 的位置，$pu_{t+j,j}$ 表

[1]为可视化展示方便，这里以字为基本单位，该图中实线空心圆表示窗口中心词，窗口中心左边用“-”号，右边用“+”号表示。

示$t+j$位置的上下文词在位置 j 时的 Embedding。

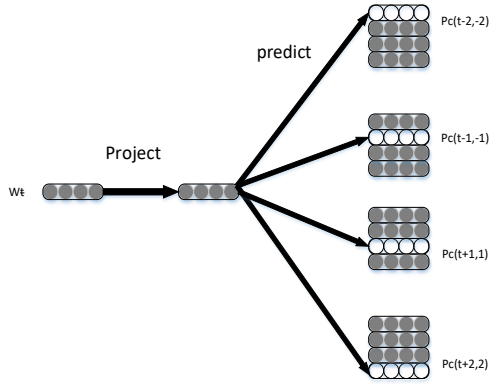

图3 位置敏感Skip-gram模型

$$p(pu \mid w_t)=\prod_{j\in[-c,0)\cup[0,c]} p(pu_{t+j,j} \mid w_t) \tag{6}$$

$$\begin{aligned}
-\log \ell &= \sum_{w\in C}\sum_{u\in Context(w)} \log(p(pu \mid w)) \\
&= \sum_{w\in C}\sum_{u\in Context(w)}\sum_{j\in[-c,0)\cup[0,c]} \log(p(pu_{t+j,j} \mid w_t)) \\
&\approx \sum_{w\in C}\sum_{u\in Context(w)}\sum_{j\in[-c,0)\cup[0,c]} \log(p(D=1 \mid pu_{t+j,j}, w_t)) \\
&+n\sum_{w\in C}\sum_{u\in Context(w)}\sum_{j\in[-c,0)\cup[0,c]} \log(p(D=0 \mid noisy_sample(pu_{t+j,j}), w_t)) \\
&= \sum_{w\in C}\sum_{u\in Context(w)}\sum_{j\in[-c,0)\cup[0,c]} \log(\sigma(\overrightarrow{pu_{t+j,j}}\cdot\overrightarrow{w_t})) \\
&+n\sum_{w\in C}\sum_{u\in Context(w)}\sum_{j\in[-c,0)\cup[0,c]} \log(\sigma(\overrightarrow{noisy_sample(pu_{i+j,j})}\cdot\overrightarrow{w_t}))
\end{aligned} \tag{7}$$

$$\overrightarrow{w_t}=\overrightarrow{w_t}-\eta*\frac{\partial(-\log \ell)}{\partial \overrightarrow{w_t}} \tag{8}$$

$$\overrightarrow{pu_{t+j,j}}=\overrightarrow{pu_{t+j,j}}-\eta*\frac{\partial(-\log \ell)}{\partial \overrightarrow{pu_{t+j,j}}} \tag{9}$$

$$\begin{aligned}
\overrightarrow{noisy_sample(pu_{t+j,j})} &= \overrightarrow{noisy_sample(pu_{t+j,j})} \\
&-\eta*\frac{\partial(-\log \ell)}{\partial \overrightarrow{(noisy_sample(pu_{t+j,j}))}}
\end{aligned} \tag{10}$$

PSG 模型的对数损失目标函数如公式(7)所示,该公式的第三步，利用负采样方法近似推导[15]。n 表示负采样样本的个数，$p(D=1 \mid pu_{t+j,j}, w_t)$ 表示 $pu_{t+j,j}, w_t$ 两词同时在训练语料中出现的概率，$p(D=0 \mid noisy\_sample(pu_{t+j,j}), w_t)$ 表示 $noisy\_sample(pu_{t+j,j})$ 和 w_t 两词不同时出现在语料中的概率。$noisy\_sample(pu_{t+j,j})$ 表示使用负采样方法对字典中的词随机采样[14]。$\overrightarrow{w_t}$，$\overrightarrow{pu_{t+j,j}}$，$\overrightarrow{noisy\_sample(pu_{t+j,j})}$ 表示 w_t，$\overrightarrow{pu_{t+j,j}}$，$\overrightarrow{noisy\_sample(pu_{t+j,j})}$ 的 Embedding。公式 8、9、10 为Embedding 更新公式，η 表示自适应学习率，随着训练样本的增加而逐渐减小[14]。

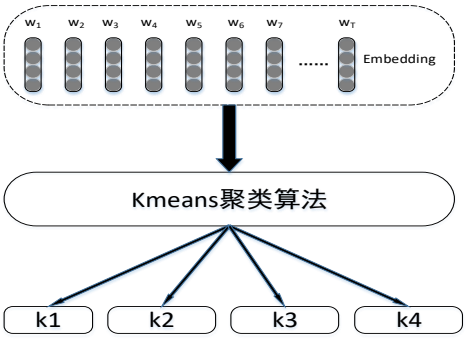

图 4 Embedding 多尺度聚类

PSG 模型训练算法流程如下：

PSG 模型训练算法

输入：词序列 $w_1, w_2, w_3 \ldots, w_T$，Embedding 维度为 d，窗口大小为 k，负采样大小为 n。

输出 $vec(w)$，$con(w, pos)$

1. 初始化：$vec(w), con(w, pos), \forall w \in W, pos \in [-k, k]$
2. for t = 1, 2, …T do
3. 取窗口中心词为 w_t
4. 上下文 $context = \{w_{t-k}, w_{t-k+1}, ..., w_{t+k-1}, w_{t+k}\}$
5. for c_t in context do
6. $cs_t' = noisy\_sample(ct, n, k)$
7. 公式 8-10 梯度更新 $vec(w_t), con(c_t, k), con(cs_t', k)$
8. end for
9. end for

上述算法中，$noisy\_sample(ct, n, k)$ 表示对上下文词 c_t，该词相对窗口中心词的位置为 k，进行负采样 n 次获取 n 个噪声词，$con(c_t, k)$ 表示 ct 相对窗口中心词位置 k 时的词向量。

2.1.3 多尺度词聚类

将 Embedding 直接用在对数线性模型中对系统的性能提高有限，因为对数线性模型更合适利用离散特征，因此需要对 Embedding 进行离散化[4]。文献[4]中使用多种方法进行离散化，实验证明对特征进行 k-means 聚类效果明显。

为了利用不同语义层次的特征，本文提出一种多尺度词聚类算法，即把 PSG 模型学习到 Embedding 利用 K-means 算法进行不同聚类尺度的词聚类。这里我们使用四种级别的聚类，分别聚 k1，k2，k3，k4 不同语义层次的词类，如图 4 所示。表 1 中的给出具体的多尺度词聚类样本，表中的数字表示词聚类的类别编号。从表 1 中可以看出，“牛”和“鸽子”在 500 和 1000 子类中，属于同一个类别，“鸽子”和“鸟”在 500、1000、5000 子类中，都属于同一个类别。这样可以利用不同语义层次的特征，提高基于 CRFs 的命名实体识别的性能。

表 1 多尺度词聚类样例

词	k1=500	k2=1000	k3=5000	k4=10000
牛	364	698	3429	8678
鸽子	364	698	2700	6203
鸟	364	698	2700	8674

2.2 条件随机场模型

条件随机场模型(CRFs)是一种给定输入节点条件下输出节点的条件概率无向图模型。本文使用的链式条件随机场，该模型广泛应用于序列标注问题中，计算给定输入序列下标记序列的分布。对应观察序列 x，对应的标注序列 y 的链式条件随机场模型的概率计算公式如下所示

$$p(y \mid x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k f_k(y_{i-1}, y_i, x)+\sum_{i,k}u_k g_k(y_i, x)\right) \tag{11}$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k f_k(y_{i-1}, y_i, x)+\sum_{i,k}u_k g_k(y_i, x)\right) \tag{12}$$

上述公式中 Z(x)是归一化因子。其中 $f_k(y_{i-1}, y_i, x)$ 表示观察序列 x 中位置 i 和 i-1 的输出节点的特征，$g_k(y_i, x)$ 表示位置 i 的输入节点和输出节点的特征，λ 和 u 表示特征函数的权重。

CRFs中定义的特征函数如下:

$$f_{y',y}(y_u,y_v,x)=\begin{cases}1 & \text{if } y_u=y',y_v=y\\0 & \text{otherwise}\end{cases} \quad (13)$$

$$g_{y,x}(y_v,x)=\begin{cases}1 & \text{if } x_v=x,y_v=y\\0 & \text{otherwise}\end{cases} \quad (14)$$

其中(y’,y)表示训练数据中的每一个状态对，(y,x)表示训练数据中的状态观察值。公式 10 表示状态转移函数，公式 11 表示特征转移函数。

本文将中文命名实体识别看作是序列标注问题，我们使用表 2 中的标签集表示文本中的实体和非实体。通过自动学习的多尺度词聚类特征，应用条件随机场模型中，来识别中文命名实体。

表 2 CRFs 的标签集

标签	标签说明
B-PER B-ORG B-LOC	实体开始标记
I-PER I-ORG I-LOC	实体中间标记
E-PER E-ORG E-LOC	实体结束标记
S-PER S-ORG S-LOC	单个词构成的实体
O	非实体标记

3 实验

本文提出基于位置敏感 Embedding 学习模型，并把学习出来的特征用于命名实体识别。因此，本文主要做了两类实验，一是比较位置敏感 Embedding 模型与传统的 Embedding 模型的表示能力。二是评估位置敏感 Embedding 对命名实体任务的影响。

本文利用公开新闻语料[2]（约 30 亿个词），首先使用 ZPar 0.7 [3]进行分词，分别基于 Skip-gram 模型和位置敏感的 Skip-gram 模型训练词的 Embedding。本实验设置窗口的大小为 5，Embedding 的维度为 200。

3.1 PSG模型评估

为了定量评估 PSG 模型是否比 SG 模型学到的 Embedding 模型具有更好的语义表示，本文使用语义资源同义词词林[4]进行评估。该词林将词语分为 12 大类，97 个中类，1400 个小类，每个小类下面若干词群，由于词群分的粒度特别细，并且词量小，为了更加有效评估，本文使用小类进行评估。

$$p@k=\frac{1}{nk}\times\sum_{w\in dict}\sum_{sw\in topk(w)}\begin{cases}1 & sw\in tongyi(w)\\0 & sw\notin tongyi(w)\end{cases} \quad (15)$$

$$sim(w_1,w_2)=\cos(e(w_1),e(w_2)) \quad (16)$$

语义相似度评估方法如公式 15 所示，n 表示同义词林中词的个数，topk(w)表示词 w 通过 Embedding 计算的语义相近的 k 个词，语义相似度用公式 16 计算，e(w)表示词 w 的 Embedding 表示。dict 表示同义词林中所有的词，tongyi(w)表示词 w 在同义词林中的所有同义词。评估结果见表 3，sg 表示 Skip-gram 模型的结果，psg 为位置敏感 Skip-gram 模型的结果。

表 3 中的两组实验进行统计有效性检验，单尾 t 检验 ($p_{value}=0.0004<0.01$)，表明 PSG 模型显著高于 SG 模型。通过表 4 实例可知，PSG 模型更加有效的学习到了词的语义信息。通过样本分析可知，PSG 模型的语义表示质量的提高主要是缓解了相关词被计算为语义相似的问题。

表 3 词义相似评估

p@k 表示 top k 的同义词准确率					
模型	p@1	p@5	p@10	p@20	p@50
SG	0.4035	0.2959	0.2460	0.2008	0.1508
PSG	0.4501	0.3365	0.2817	0.2312	0.1743

表 4 语义相似词

模型	词	语义最近的 5 个词
sg	武大	樱花；厦大；磨山；樱照；赏樱
	电影	动画；文艺片；上映；好莱坞；新片
psg	武大	厦大；南大；珞珈山；北大；华农
	电影	影片；动作片；喜剧片；商业片；文艺片

3.2 命名实体识别评估

命名实体识别实验使用 98 年 1 月份《人民日报》标注语料[5]，识别其中的人名、地名、机构名。该语料共包含 19484 个句子。为了更加客观的评估不同模型的性能，我们采取十折交叉验证进行实验。

表 5 命名实体识别特征模板

特征名称	特征模板
词特征	$w_t:t\in[-2,2];w_tw_{t+1}:t\in[-1,1]$
词聚类特征	$c_t:t\in[-2,2];c_tc_{t+1}:t\in[-2,1];$ $c_tc_{t+1}c_{t+2}:t\in[-2,0]$

本文使用 CRFs 模型进行命名实体识别，使用的特征模板如表 5 所示，特征模板表示上下文原始词特征，词聚类特征表示词的 Embedding 聚类特征。符号 w_t 表示句子中的第 t 个词，w_tw_{t+1} 表示二元词。c_t 表示 w_t 对应的 Embedding 的聚类特征，以此类推。

本文把只使用词特征的系统作为基线系统(base)，分别使用 SG 模型的 Embedding 聚 1,000 类和 10,000 类的特征系统(sg_1k,sg_10k)，使用多尺度词聚类特征的系统为(sg_scale)。系统 psg_1k,psg_10k 使用 PSG 模型的 Embedding，psg_scale 使用 PSG 模型 Embedding 的多尺度离散化特征。sg_scale 和 psg_scale 使用的多尺度词聚类的控制参数 k1=500,k2=1,000,k3=5,000,k4=10,000。本文把十折交叉验证样本集的平均值作为最终的命名实体识别结果，如表 6 所示。从表 6 可知，sg_1k 系统比 base 系统 F 值提高了 1.9%，召回率提高了 3.72%，但是准确率降低了 0.17%，该结果表明 Embedding 模型学习到的特征可以提高命名实体识别的性能，主要提高了模型的泛化能力，但是降低了模型的准确度。sg_10k 比 sg_1k 的 F 值提高了 0.21%，准确率也有所提高，表明细粒度的聚类有助于降低泛化风险。sg_scale 相对 sg_10k 的提高，因利用不同层次的语义聚类特征，使得泛化能力进一步提升。

表 6 命名实体识别结果

系统	准确率	召回率	F 值
base	96.16	87.67	91.72
sg_1k	95.99	91.39	93.63
sg_10k	96.39	91.43	93.84
sg_scale	96.1	92.21	94.11
psg_1k	96.1	92.19	94.1
psg_10k	**96.52**	92.12	94.27
psg_scale	96.14	**93.04**	**94.57**
pas_scale	96.12	92.39	94.22

对比 PSG 模型和 SG 模型在不同参数设置条件下的命名实体识别性能，可以发现 PSG 模型在识别性能的准确率和

[2] http://pullword.com/
[3] http://sourceforge.net/projects/zpar/files/0.7/
[4] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm
[5] http://www.icl.pku.edu.cn/icl_res/

召回率上都要优于 SG 模型，在两个模型都在最优的实验条件下，基于 PSG 模型的系统比基于 SG 模型的系统 F 值提高了 0.46%，召回率提高了 0.83%，性能的提高主要依赖于模型泛化能力的提高。

通过表 5 中例子，在命名实体识别评估中，PSG 模型优于 SG 模型，主要得益于缓解了语义表示偏差，例如“武大”是一个学校名称，属于命名实体定义中的机构，利用 SG 模型学习到的语义表示，“武大”最相似的词是“樱花”，但是“樱花”不是命名实体，这给条件随机场模型训练带来了误差。而 PSG 模型中，“武大”最相似的词中有四个都是学校名称，由此可见基于 PSG 模型的命名实体识别系统的提高主要是因为 PSG 模型学习到的语义特征更加有效，具有更强的泛化能力。

此外，本文对比了文献[16]的 PAS Skip-gram 模型，该模型也利用位置信息增强 Embedding 的语义表示质量，本文把学习到的 Embedding，使用该模型的 Embedding 多尺度词聚类特征系统为 pas_scale.该系统的 F 值比 psg_scale 模型低 0.35%，比 sg_scale 提高 0.11%。说明 PAS Skip-gram 模型能够提高中文命名实体识别的性能，但是通过分析实际的例子，例如在 PAS Skip-gram 模型中，“武大”最相似的词是“樱照”，说明该模型提高了 Embedding 语义表示质量的时候，并不能缓解相关的词语义相近的问题。

3.3 实验总结

本文基于位置敏感 Embedding 模型和多尺度聚类的命名识别系统比基线系统(base)F 值提高了 2.85%，在基本上不降低准确率的基础上，召回率提高了 5.73%。比只利用 Embedding 模型的命名识别系统(sg_10k)提高 0.73%。经过统计有效验证，系统 psg_scale 比 base 和 sg_10k 都有显著性提高。该实验结果证明本文提出的基于位置敏感 Embedding 模型的中文命名实体识别方法的有效性。

4 结束语

本文提出了一种基于位置敏感 Embedding 的中文命名实体识别方法，一方面通过位置敏感 Embedding 模型自动化学习词的语义表示，降低特征的稀疏性，减小命名实体识别对手工特征和语言学知识依赖，另一方面提高了 Embedding 的语义表示质量，减小模型的泛化误差，通过多尺度词聚类算法，尽可能利用不同层次的语义特征，提高中文命名实体识别的性能。

为了更加有效地利用 Embedding 学到的特征，下一步工作是利用深度学习模型例如循环神经网络条件随机场[16]。直接将 Embedding 输入到神经网络中进一步提高中文命名实体识别的性能。

参考文献

[1] 冯元勇，孙乐，张大鲲，等. 基于小规模尾字特征的中文命名实体识别研究[J]. 电子学报，2008，36(9)：1833-1838.
[2] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[3] Guo J, Che W, Wang H, et al. Learning sense-specific word embeddings by exploiting bilingual resources[C]//Proceedings of COLING. 2014: 497-507.
[4] Guo J, Che W, Wang H, et al. Revisiting embedding features for simple semi-supervised learning[C]//Proceedings of EMNLP. 2014: 110-120.
[5] Sundheim B M, Chinchor N. Named entity task definition, version 2.1[C]//Proceedings of the sixth message understanding conference. 1995: 319-332.
[6] 王宁，黄锦辉. 中文金融新闻中公司名的识别[J]. 中文信息学报，2002，16(2)：1-6.
[7] 罗智勇，宋柔. 一种基于可信度的人名识别方法[J]. 中文信息学报，2005，19(3)：67-72.作者. 文章名[C]//会议名. 会议召开城市，国家：出版社，年份.
[8] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]//proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 473-480.
[9] 张玥杰，徐智婷，薛向阳. 融合多特征的最大熵汉语命名实体识别模型[J]. 计算机研究与发展，2008，45(6).
[10] 彭春艳，张晖，包玲玉，等. 基于条件随机域的生物命名实体识别[J]. 计算机工程，2009，35(22)：197-199.
[11] 邱莎，王付艳，申浩如，等. 基于含边界词性特征的中文命名实体识别[J]. 计算机工程，2012，38(13)：128-130.
[12] 向晓雯，史晓东，曾华琳. 一个统计与规则相结合的中文命名实体识别系统[J]. 计算机应用，2005，25(10)：2404-2406.
[13] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.
[14] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[15] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[C]//Advances in Neural Information Processing Systems. 2014: 2177-2185.
[16] Qiu L, Cao Y, Nie Z, et al. Learning Word Representation Considering Proximity and Ambiguity[C]//AAAI. 2014: 1572-1578.
[17] Yao K, Peng B, Zweig G, et al. Recurrent conditional random field for language understanding[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 4077-4081.