基于自动编码器的中文词汇特征无监督学习*

张开旭，周昌乐

（厦门大学信息科学与技术学院，福建省 厦门市 361005）

摘要：大规模未标注语料中蕴含了丰富的词汇信息，有助于提高中文分词词性标注模型效果。本文从未标注语料中抽取词汇的分布信息，表示为高维向量。进一步使用自动编码器神经网络，无监督地学习对高维向量的编码算法，最终得到可直接用于分词词性标注模型的低维特征表示。在宾州中文树库 5 数据集上的实验表明，所得到的词汇特征对分词词性标注模型效果有较大帮助，在词性标注上优于主成份分析与 *k* 均值聚类结合的无监督特征学习方法。

关键词： 无监督特征学习；中文分词；词性标注

中图分类号：TP391 **文献标识码：**A

Unsupervised Feature Learning for Chinese Lexicon Based on Auto-Encoder

Kaixu Zhang, Changle Zhou

(Xiamen University, Xiamen, Fujian 361005, China)

Abstract: Large-scale unlabeled data contains abundant lexical information for NLP tasks such as Chinese word segmentation and POS tagging. This work extracted high-dimensional distributional lexical information from a large-scale unlabeled Chinese corpus. An auto-encoder then performed the unsupervised dimension reduction. The learned low-dimensional lexicon features were used as new lexical features for a joint Chinese word segmentation and POS tagging task. Experiments on the Chinese Treebank 5 corpus showed that the additional lexicon features improve the performance and are better than those features learned by using the principal component analysis and the *k*-means algorithm.

Key words: unsupervised feature learning；Chinese word segmentation；part-of-speech tagging

1 引言

中文分词词性标注是中文自然语言处理的重要任务。训练分词词性标注模型依赖于人工标注的训练数据。然而人工标注的数据规模有限，难以涵盖多样的文本和广泛的词汇，这制约了分词词性标注模型的性能。

从未经标注的语料中获取有用的词汇信息，提高中文分词词性标注的效果，一方面避免了高成本的人工标注，另一方面利用更加广泛的词汇信息可以克服人工标注数据的局限性。本文试图从大规模的未标注语料中提取出中文词汇的分布信息，并经过自动学习得到相应的特征以提高现有分词词性标注模型效果。其中最大的挑战在于如何将词汇的高维分布信息转化为低维特征表示。

一个词的上下文分布多种多样，如果一一统计，不但存在大量冗余或无用信息，而且对于大规模数据在时间、空间复杂度上也不可接受。本文通过启发式的方法，选出最能代表词汇特征的上下文，只统计词汇在这些特定上下文中的出现情况，用以代表词汇的分布信息，解决了在大规模语料中抽取分布信息的时间、空间复杂度问题。

直接将此分布信息作为特征，其维度仍然过高。本文进一步使用自动编码器

* **定稿日期：**2013 年 7 月 15 日

基金项目：国家自然科学基金（61273338）；教育部高等学校博士学科点专项科研基金(新教师类)（20120121120046）；福建省自然科学基金（2010J01351）

作者简介：张开旭（1984——），男，博士后，主要研究方向为中文自然语言处理；周昌乐（1959——），男，教授，主要研究方向为人工智能。

（auto-encoders），无监督地学习对这一高维分布信息的编码函数，从而得到相应的低维表示，可直接用作分词词性标注模型的特征。自动编码器是一种无监督学习高维数据的低维表示的神经网络，在深度学习中被广泛用于无监督特征学习，并在图像分类等任务中表现出了很好的效果[1,2]。

实验中，我们自动选择了 1346 个上下文，在 260 亿汉字的互联网语料中统计了 52876 个多字词的分布信息，使用自动编码器进行特征无监督学习，最终对每个词得到 50 维的词汇低维稀疏表示。在宾州中文树库 5.0 数据集上的实验表明该方法得到的词汇特征对分词词性标注模型的效果有较大提升，在词性标注上优于主成份分析与 k 均值聚类结合的方法。此外在训练自动编码器时对输入引入噪音，得到的降噪自动编码器也能产生更高质量的特征。

2 相关工作

基于人工标注训练集的中文分词词性标注模型已经被广泛深入地研究，其中包括基于字标注的模型[3]，基于亚词的模型[4]，基于词的模型[5]等，本文所用分词词性标注基线方法改进自基于词图的模型[6]。

相关工作也集中在使用标注训练集之外的未标注语料提高效果。其中包括使用邻接变化数[7]等用于提高中文分词效果的统计量。此外也有基于无标注数据自动分析结果的特征，Wang 等人[8]和 Sun 等人[9]等人将未标注的 Gigaword 语料进行自动分析，从其结果中提炼特征帮助分词词性标注。但是以上方法均难以直接用于本文中的大规模语料。

除了使用未标注语料，Jiang 等人[10,11]和 Sun 等人[12]也尝试利用具有不同标注规范的额外语料。他们均使用北京大学标注的人民日报语料库，来提高宾州中文树库 5.0 数据集上的分词模型或分词词性标注模型的效果。

自动编码器被广泛用于深度学习中的无监督特征学习。Coates 等人[1]使用基于支持向量机的图片分类任务，比较了若干无监督特征学习算法，除稀疏自动编码器外效果最好的是主成份分析与 k 均值聚类结合的方法，因此本文也在实验中对这两种方法加以比较。此外降噪自动编码器[2]也被用以提高深度学习分类器的效果，本文也将研究其是否对中文词汇特征自动学习有效。

词嵌入（word embedding）是通过建立神经网络语言模型，对词进行低维稠密连续表示[13]。但在大规模语料上学习神经网络语言模型的速度较慢，因此在本实验中未采用这种方式进行特征学习。

3 词汇信息获取

3．1 基于分布的词汇信息表示

本节将讨论使用一个高维布尔向量反映词汇的上下文分布信息。向量的每一维，表示一种上下文，目标词所对应的向量某一维为 1，表示目标词能够出现在相应的上下文中，反之则表示该词不能出现在相应的上下文中。我们期望这样的向量，能够尽可能的反映目标词的句法、语义信息，以便在具体的应用中加以利用。

要在大规模语料上实现这一目标，需要解决两个问题：首先，给定上下文和目标词后，如何判断目标词能够出现在相应上下文中，即向量分量的计算问题；其次，词汇出现的上下文是无穷无尽的，应该选择哪些上下文作来刻画词汇的分布信息，即向量维度的选择问题。后两小节分别对这两个问题进行讨论。

3．2 分量计算

如果目标词与特定上下文的共现概率大于某一阈值，就可将目标词布尔向量相应维度的分量置为 1。但由于存在噪音以及在目标词出现频率较小的情况下，使用最大似然估计得到

的共现概率不太准确，在此以如下方式确定分量取值。

设目标词为 w，在语料库中出现的频次为 n，设词串与上下文 c 在语料库中共同出现的次数为 m，如果

$$m \geq np + R\sqrt{np(1-p)} \quad (1)$$

则称目标词 w 与上下文 c 匹配。本公式基于二项分布假设检验，p 为零假设时目标词与上下文 c 的共现概率。在本实验中，使用单边 95%的置信区间，即 R=1.67，并且令零假设为一个较小的概率值 p=10^{-4}。与最大似然估计使用的公式 $m \geq np$ 相比，本实验所用公式的不等式右边增加了一项，更为保守，可以减少可能的噪音。并且在目标词出现频次 n 趋于无穷大时，两种方法等价。

3．3 维度选择

本小节讨论如何确定上下文的集合用以统计词汇的分布信息。

为了使同一个上下文所匹配上的词的句法、语义更为单一，更具特异性，我们使用目标词左边、右边出现的两个词组成的词对$\langle w_1,w_2\rangle$来表示上下文 c。例如句子片段“材料 利用率 高”中，目标词“利用率”就与上下文“$\langle$材料,高$\rangle$”共现。

根据以上方式定义的上下文数量相当庞大，不可能为所有目标词一一统计可能的上下文。因此需要确定一个上下文的子集用以统计词汇的分布信息。我们主要排除那些匹配的词不多或者过于特殊不具有句法、语义意义的上下文。在此我们的假设是，如果一个上下文所匹配的词有句法、语义意义，那么应该有其它的上下文也能正好匹配上这些词。例如，如果我们发现“$\langle$材料,高$\rangle$”、“$\langle$材料,低$\rangle$”、“$\langle$物资,高$\rangle$”等一系列上下文所能匹配的词较为一致，就说明它们所匹配的词在句法、语义上有一定共现，并且说明这些上下文能够反映词汇的某些句法、语义性质，是比较有效的上下文。

基于以上假设，设计以下启发式方法确定上下文集合。设上下文词对 c_1 和 c_2，分别能够匹配的词的集合为 W_1 和 W_2。则定义 c_1 与 c_2 的相似度为它们能匹配上的词的集合之间的 Jaccard 距离

$$\mathrm{Sim}(c_1,c_2) = \mathrm{Jaccard}(W_1,W_2) = |W_1 \cap W_2| \,/\, |W_1 \cup W_2| \quad (2)$$

根据得到的相似度矩阵，使用吸引力传播（Affinity Propagation）聚类算法，对词对聚类。得到类别成员数大于 5 的类别所包含的所有上下文词对，用作表示词汇信息向量的分量。

4 无监督特征学习

基于分布的词汇信息维度仍然较高，不适合直接用于分词词性标注等任务。本节讨论如何由高维的词汇信息，不根据具体任务或者具体任务的正确标注，无监督地学习出低维的可用于具体任务的特征向量。本文主要讨论在深度学习中使用较多的自动编码器以及基于主成份分析和 k 均值聚类的方法。

4．1 自动编码器

自动编码器是一种多层前传神经网络，可以用来对高维数据降维，得到低维的特征向量，其在深度学习中被广泛运用。

在将自动编码器用于无监督特征学习时，通常使用有一个输入层、一个隐层以及一个输出层的神经网络。设输入样本的向量表示为 x，通过以下方式可得到隐层和输出层的激活情况：

$$\mathbf{y}=S(\mathbf{W}\mathbf{x}+\mathbf{b}) \quad (3)$$

$$\mathbf{z}=S(\mathbf{W}^{\mathrm{T}}\mathbf{y}+\mathbf{b}') \quad (4)$$

$\langle q_0.\text{conf}\rangle, \langle q_0.\text{conf}, s_0.\text{conf}\rangle$
$\langle q_0.\text{w}\rangle, \langle q_0.\text{t}\rangle, \langle q_0.\text{len}\rangle, \langle q_0.\text{w}, q_0.\text{t}\rangle, \langle q_0.\text{t}, q_0.\text{len}\rangle$
$\langle q_0.\text{w}, s_0.\text{w}\rangle, \langle q_0.\text{w}, s_0.\text{t}\rangle, \langle q_0.\text{t}, s_0.\text{w}\rangle, \langle q_0.\text{t}, s_0.\text{t}\rangle$
$\langle q_0.\text{len}, s_0.\text{w}\rangle, \langle q_0.\text{w}, s_0.\text{len}\rangle, \langle q_0.\text{len}, s_0.\text{t}\rangle, \langle q_0.\text{len}, s_0.\text{len}\rangle$
$\langle q_0.\text{w}, s_0.\text{t}, s_0.\text{t}\rangle, \langle q_0.\text{w}, q_0.\text{t}, s_0.\text{w}\rangle, \langle q_0.\text{w}, s_0.\text{w}, s_0.\text{t}\rangle$
$\langle q_0.\text{t}, s_1.\text{t}\rangle, \langle q_0.\text{t}, s_0.\text{t}, s_1.\text{t}\rangle, \langle q_0.\text{len}, s_0.\text{len}, s_1.\text{len}\rangle$

表 1 基线分词词性标注模型的特征模板

其中 $S(x)=1/(1-e^{-x})$为 Sigmoid 函数，并注意与一般的多层前传神经网络不同的是，前后两个公式使用的权重矩阵 $\mathbf{W}$、$\mathbf{W}^{\mathrm{T}}$ 互为转置。

与有监督学习不同，自动编码器的学习目标是使输出层尽量还原输入层的状态，既使得 **z** 尽量与 **x** 相同，本文使用损失函数

$$L(\mathrm{X},\mathrm{Z})=\sum_{i=1}^{n}\mathrm{KL}(\mathbf{x}_i \parallel \mathbf{z}_i) \tag{5}$$

这里我们采用了矩阵表示，**X** 是一个由 n 个样本的向量组成的矩阵，$\mathrm{KL}(\mathbf{x}_i\|\mathbf{z}_i)$，是输入向量 $\mathbf{x}_i$ 与输出向量 $\mathbf{z}_i$ 的 KL 散度，用以度量它们之间的区别。

而由于隐层 **y** 的维度比 **x** 的维度小得多，所以隐层 **y** 可以学习到输入样本的低维表示，并且能够通过解码尽量包含与高维表示相同的信息。使用没有标注的数据集 **X** 进行自动编码器神经网络的学习。最后对于任何输入向量 **x**，计算其对应的隐层向量 **y**，就得到了输入向量的一个低维编码。

自动编码器权重的训练采用随机梯度下降算法，使用以下公式更新权重矩阵

$$\mathbb{W}\leftarrow\mathbb{W}-\eta\frac{\partial L(\mathrm{X},\mathrm{Z})}{\partial\mathbb{W}} \tag{6}$$

其中 η 为更新的步长，参数 **b** 和 **b**’采用相同方式更新。

4．2 降噪自动编码器

为了使自动编码器更具鲁棒性，更好地对存在噪音的数据进行编码，在训练时对输入进行污染[2]，即使用 **X**’代替 **X** 作为输入，**X**’的元素定义为

$$\mathbf{X}'_{i,j}=\mathbf{X}_{i,j}\mathbf{B}_{i,j} \tag{7}$$

其中矩阵 **B** 的元素独立采样自一个伯努利分布。但在训练自动编码器时仍然要求自动编码器的输出 **Z** 尽量还原真实的输入 **X** 而非 **X**’。

4．3 稀疏自动编码器

在训练自动编码器时，通常还可以引入额外的约束，限制隐层神经元的激活数目，使得对于一个样本，只有少部分隐层神经元被激活，这也是对数据进行稀疏编码（sparse coding）的实现方式之一。稀疏编码被证明有可能提高模型效果[1]。

在本实验中，在损失函数中引入对隐层神经元激活数目的约束项

$$L(\mathrm{X},\mathrm{Z})=\sum_{i=1}^{n}\mathrm{KL}(\mathbf{x}_i \parallel \mathbf{z}_i)+\beta\sum_{j=1}^{b}\mathrm{KL}(\rho \parallel \rho_j) \tag{8}$$

其中 $\rho_j=\sum_{i=1}^{n}Y_{i,j}/n$ 为第 j 个隐藏神经元的激活率，约束项的意义是使得隐层每一个神经元的实际激活率接近设定的一个较小的值 ρ。注意公式中第一个求和项是对 n 个样本求和，

而第二个求和项是对 b 个隐层神经元求和。

4．4 使用主成份分析和 k 均值聚类的特征学习方法

除了使用自动编码器，本文也使用基于主成份分析（PCA, Principal Component Analysis）和 k 均值聚类的方法进行无监督特征学习。

首先对输入数据进行主成份分析降维，并进行白化（whiten）。经过 PCA 白化后，表示数据样本的向量每一维的均值为 0，方差为 1 并且相互独立，即其协方差矩阵为单位阵。

进一步地，使用 k 近邻聚类方法对样本进行聚类，为了避免硬聚类造成的噪音，对每个样本，取前 h 个最近邻作为样本的特征表示。

5 提升分词词性标注效果

5．1 基线模型

本文使用基于词图的中文分词词性标注联合模型。输入是由基于字标注的分词词性标注模型输出的词图，基于词图的模型在输入的词图中找出最优路径作为最终的分词词性标注结果。

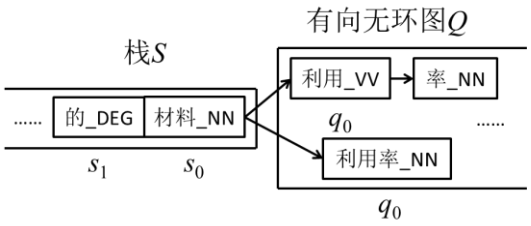

图 1 基于词图的分词词性标注模型的解码过程

如图 1 所示，将最优路径的选择过程看作从一个有向无环的词图中依次找出元素压栈的过程。最后当 Q 处理完毕，栈中的元素就构成最终的分词词性标注结果。由于在一个位置能够压栈的元素可能有多种选择，模型根据当前栈顶、次栈顶元素和待压栈元素判断压栈动作的分数，最后在所有可能的压栈动作序列中搜索出动作分数之和最高的作为最终的输出。

本文采用 Huang 等人[14]的方法，进行解码和模型参数的学习。所使用的特征模板列在表 1 中。其中 s_1、s_0 分别是次栈顶和栈顶，q_0 是待压栈的词，s_0.w 表示该词本身，s_0.len 表示的词长，s_0.t 表示词的词性，s_0.conf 表示该词在词图上的置信度。

5．2 无监督特征和其它特征的引入

在获取词汇信息时，我们在大规模语料中统计了字符串能够匹配上的不同上下文模板的情况，构成词的字符串能够匹配的不同上下文较多，而不是词的字符串能够匹配的上下文较少。根据字符串匹配模板多少，我们可以设计新的受限邻接变化数[15]特征，既表 2 中的 rav，表示该词匹配上的上下文的个数。

进一步，我们引入无监督特征学习得到的特征，相关的特征模板见表 2 中的 ae 特征和 k 特征。其中 ae_j 表示该词在自动编码器编码之后的向量第 j 维分量的值大于 0.9。类似的，可使用 k 均值方法引入无监督特征，其中 k_j 表示该词最近的 5 个类中心中有 j。

最后，与 Jiang 等人[10]、Sun 等人[12]的方法类似，使用北大人民日报半年语料库训练一个基于字标注的分词词性标注模型，对 CTB5 语料进行分词词性标注，用标注结果作为新的特征，见表 2 中的 pku。其中 pku 表示该词使用人民日报语料训练的模型中被标注的词性。

RAV 特征模板	$\langle q_0.\text{rav}\rangle$, $\langle q_0.\text{rav}, s_0.\text{rav}\rangle$, $\langle q_0.\text{rav}, q_0.\text{w}\rangle$ $\langle s_0.\text{rav}, q_0.\text{w}\rangle$, $\langle q_0.\text{rav}, s_0.\text{w}\rangle$
自动编码器特征模板	$\langle q_0.\text{ae}_i, q_0.\text{w}\rangle$, $\langle q_0.\text{ae}_i, s_0.\text{w}\rangle$, $\langle s_0.\text{ae}_i, q_0.\text{w}\rangle$
k 均值聚类特征模板	$\langle q_0.\text{k}_i, q_0.\text{w}\rangle$, $\langle q_0.\text{k}_i, s_0.\text{w}\rangle$, $\langle s_0.\text{k}_i, q_0.\text{w}\rangle$
PKU 特征模板	$\langle q_0.\text{t}, q_0.\text{pku}\rangle$, $\langle s_0.\text{t}, q_0.\text{pku}\rangle$, $\langle q_0.\text{t}, s_0.\text{pku}\rangle$ $\langle s_0.\text{w}, q_0.\text{pku}\rangle$, $\langle q_0.\text{w}, s_0.\text{pku}\rangle$

表 2 额外的特征模板

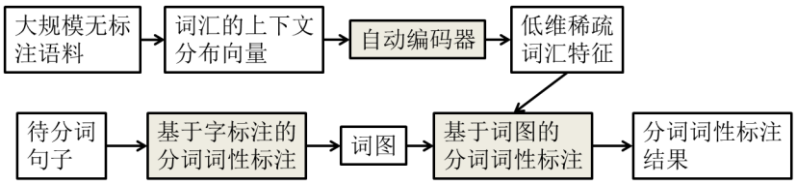

图 2 本文所用分词词性标注模型流程图

最终，我们使用的分词词性标注系统流程图以及所添加的无监督词汇特征如图 2 所示。

6 实验

6．1 实验设置

本文用以抽取中文词汇信息的资源为 SogouT 互联网语料[1]2008 版。其中包含了 HTML 格式的网页文件，对内容进行自动筛选，找到其中的正文部分，得到共 6800 万网页、260 亿字符的数据集。

为了对语料进行分词，使用的是一个基于字标注的中文分词模型，其训练语料来自中文树库 CTB5，按照 Zhang 等人[5]的划分方式划分训练集、开发集和测试集，用以训练分词模型的为其中的训练集。

同时，为了检验无监督词汇特征学习得到的词汇特征的有效性，仍然使用在 CTB5 语料上进行的中文分词词性标注任务。

与 Jiang 等人[10]、Sun 等人[12]的工作类似，本文会使用在人民日报训练分词词性标注模型，使用其对 CTB5 语料进行分词词性标注，将所得到的结果作为额外的特征。本文所使用的人民日报标注语料库[2]与原版略有不同，比如已经将汉族人名的姓氏和名字两部分进行了合并。

在大规模语料中抽取词的分布信息，需要分词模型进行分词，而基于词图的分词词性标注基线模型也需要分词词性标注模型生成词图，这两个模型均使用 THULAC 中文词法分析工具包[3]在 CTB5 的训练集上进行训练得到。分词和词性标注的效果使用 F 值进行评测[3]。

6．2 词汇信息获取

首先使用在 CTB5 训练语料上训练的中文分词模型对 SogouT 中抽取的文本进行中文分词，得到一个高频词表，同时得到词表中的词与各种上下文词对的匹配情况。根据 3.3 节的方法定义上下文词对的相似度，使用吸引力传播聚类工具对上下文词对聚类，选择其中类别成员大于 5 的共 131 个类别中的 1346 个上下文词对，作为词汇向量空间表示中的维度。

[1] http://www.sogou.com/labs/dl/t.html
[2] http://vdisk.weibo.com/s/8Viac
[3] http://nlp.csai.tsinghua.edu.cn/thulac/

⟨去,参加⟩⟨飞往,，⟩⟨届,国际⟩⟨从,引进⟩⟨在,出差⟩⟨在,警方⟩
⟨那么,。⟩⟨蛮,的⟩⟨非常,！⟩⟨很,地⟩⟨挺,的⟩⟨很,很⟩
⟨我们,了⟩ 他们,了⟩⟨没有,过⟩⟨次,了⟩⟨地,了⟩⟨已,了⟩⟨上,了⟩
⟨名,说⟩⟨位,表示⟩⟨位,向⟩⟨位,认为⟩⟨位,这样⟩
⟨有,名⟩⟨在,个⟩⟨了,次⟩⟨有,条⟩⟨了,架⟩⟨了,块⟩

表 3 使用吸引力传播算法对上下文进行聚类的部分结果

表 3 是聚类结果中的部分类别及其代表成员，每一行为同一类别中的不同上下文。可以看到，一方面所选出的上下文词对能够匹配大量的词汇，另一方面，每一个类别中的上下文词对可以体现一个词汇的相同的句法、语义特征。因此用这些上下文词对的匹配情况来表征一个词的句法、语义信息，是有效的。同时由于对于每个词仅需要记录其在一千余种不同上下文中的出现情况，其计算的时间、空间复杂度也适合进行大规模数据的处理。

最终，使用这 1346 个上下文词对，在 260 亿字符的数据集上对 52876 个高频多字词的分布信息进行了统计。

6．3 词汇特征无监督学习

根据 4.1 节的方法，使用自动编码器对抽取出的词汇信息进行进一步的编码。自动编码器的人工神经网络的可见层有 1346 个神经元，隐层有 50 个神经元。为给输入引入噪音，每次迭代的时候随机选择每个输入数据 10%的维度将其置为 0。为实现稀疏自动编码，将公式 6 中的 ρ 置为 0.1，β 置为 1。训练采用随机梯度下降算法迭代 15 次。通过自动编码器的编码，最终对于每个词，得到了一个 50 维的零一向量，作为其词汇特征向量。

权重较大的分量	激活程度较高的样本
⟨以,的⟩⟨能,的⟩⟨可以,的⟩⟨又,地⟩	灵活 完整 明智 乐观 精确 轻松 准确
⟨我,了⟩⟨是,了⟩⟨他,了⟩⟨她,了⟩	呼唤 宣示 袭击 享受 断定 决裂 爆发
⟨，,要⟩⟨。,是⟩⟨的,要⟩⟨的,有⟩	师傅 大老板 女生 小伙子 女人 MM 哥哥
⟨所,的⟩⟨通过,的⟩⟨你,的⟩⟨从,的⟩	开辟 设立 构筑 营造 选用 搭建 创建

表 4 使用自动编码器学习后部分隐层单元学习结果

表 4 展示了 50 个隐层神经元之中的 4 个的相关信息，第一列表示能最大程度激活该神经元的部分上下文模板，第二列列举了能最大程度激活该神经元的部分词汇。可见它们也都有一定的句法、语义共性。

为了与自动编码器特征学习效果进行对比，在使用 k 均值聚类的方法无监督学习特征时，对经过 PCA 白化的数据使用 k 均值聚出 50 个类别。并且对于任意一个样本，选择与其距离最近的 5 个类中心作为其特征表示。

6．4 提升分词词性标注性能

表 5 是在 CTB5 语料上分词词性标注任务中的结果。结果中标有“$^{(*)}$”的表示方法额外使用了人民日报标注数据集。

在相关工作中，Wang 等人[8]使用经过自动分析的 Gigaword 新闻语料中提取的提升帮助 CTB5 语料上的分词词性标注效果，Jiang 等人[10,11]和 Sun 等人[12]的方法使用北大人民日报语料库提升 CTB5 语料上的分词词性标注效果。

本文基线模型能达到较好的效果，引入 RAV 特征可以小幅提高分词的效果，但对词性标注的帮助不大。

再考察无监督词汇特征学习得到的特征。不论是引入 k 均值聚类的特征还是自动编码器

方法	分词 F 值	词性标注 F 值
Wang_2011[8]	0.9811	0.9418
(*)Jiang_2009[10]	0.9823	0.9403
(*)Jiang_2012[11]	0.9843	无
(*)Sun_2012[12]	未提供	0.9467
基线	0.9796	0.9387
基线+RAV	0.9808	0.9386
基线+RAV+*k* 均值聚类特征	0.9828	0.9420
(*)加入 PKU 模型解码结果特征	**0.9845**	0.9464
基线+RAV+自动编码特征	0.9834	0.9448
特征学习无稀疏编码约束	0.9833	0.9447
特征学习不引入噪音	0.9819	0.9432
(*)加入 PKU 模型解码结果特征	0.9843	**0.9480**

表 5 在 CTB5 数据集上本文各种方法效果与已有方法比较方，方法名称中标注有“(*)”的表示其使用了额外的人工标注语料。

的特征，分词词性标注的效果均有明显提升，在词性标注上自动编码器的效果较优。可见在大规模语料中无监督学习得到的特征对分词词性标注模型有很大的帮助。对自动编码器而言，系数编码约束的映入对效果影响不大，而训练时噪音的引入对无监督特征学习的质量是有帮助的。

最后，进一步加入人民日报语料库相关特征后，模型效果有了进一步提高。无论是否使用额外的标注数据集，本文方法的效果均超过了相关工作。

7 结论

本文研究利用自动编码器进行中文词汇特征的无监督学习。首先从大规模无标注语料中抽取词汇的高维分布信息。并使用自动编码器无监督学习得到的低维特征提升中文分词词性标注任务效果，以检验所得到特征的有效性。此外，还是用主成份分析与 *k* 均值聚类的方法进行无监督特征学习，与使用自动编码器的方法进行对比。在宾州中文树库 5 数据集上的实验表明，从大规模无标注语料中学习的词汇特征显著提升了分词词性标注任务的效果，使用自动编码器的方法要优于 *k* 均值聚类的方法。并且本文模型的效果超过了所有相关工作。

本文所提出的特征，将词汇分布式地表示为一个低维向量，与直接使用词汇本身作为特征相比，特征数目非常少，可以避免由于特征过多造成的数据稀疏问题，并且对训练集中未出现的词也能够更好的处理。这应是其能够提高现有分词词性标注模型性能的原因之一。

本文只是无监督特征学习在中文信息处理中的一次尝试。未来将研究更好的词汇分布信息的抽取、表示方法，以及汉字、字串分布信息的抽取、表示方法。进一步地，可通过自动编码器学习词串、句法关系等更大粒度语言单位的特征表示。并在深度神经网络的框架下进行统一的参数学习，用于分词、词性标注、句法分析等多种自然语言处理任务。

参考文献：

[1] Coates, Adam and Ng, Andrew Y and Lee, Honglak. An analysis of single-layer networks in unsupervised feature learning[C]// International Conference on Artificial Intelligence and Statistics. 2011, 215—223.

[2] Vincent, Pascal and Larochelle, Hugo and Bengio, Yoshua and Manzagol, Pierre-Antoine. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th international conference on Machine learning. 2008, 1096—1103.

[3] Ng, Hwee Tou and Low, Jin Kiat. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once?

Word-Based or Character-Based?[C]//Proceedings of EMNLP 2004. Barcelona, Spain: Association for Computational Linguistics. 2004, 277–284.

[4] Sun, Weiwei. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics. 2011, 1385–1394.

[5] Zhang, Yue and Clark, Stephen. Joint Word Segmentation and POS Tagging Using a Single Perceptron[C]// Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics. 2008, 888–896.

[6] Jiang, Wenbin and Mi, Haitao and Liu, Qun. Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging[C]// Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, UK: Coling 2008 Organizing Committee. 20008, 385–392.

[7] Feng, Haodi and Chen, Kang and Kit, Chunyu and Deng, Xiaotie. Unsupervised Segmentation of Chinese Corpus Using Accessor Variety[C]// Natural Language Processing IJCNLP. 2005, 694–703.

[8] Wang, Yiou and Jun'ichi Kazama, Yoshimasa Tsuruoka and Tsuruoka, Yoshimasa and Chen, Wenliang and Zhang, Yujie and Torisawa, Kentaro. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data[C]// Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. 2011, 309–317.

[9] Sun, Weiwei and Uszkoreit, Hans. Capturing Paradigmatic and Syntagmatic Lexical Relations: Towards Accurate Chinese Part-of-Speech Tagging[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics. 2012, 242–252.

[10] Jiang, Wenbin and Huang, Liang and Liu, Qun. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study[C]// Proceedings of the 47th ACL. Suntec, Singapore: Association for Computational Linguistics. 2009, 522–530.

[11] Jiang, Wenbin and Meng, Fandong and Liu, Qun and Lü, Yajuan. Iterative Annotation Transformation with Predict-Self Reestimation for Chinese Word Segmentation[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics. 2012, 412–420.

[12] Sun, Weiwei and Wan, Xiaojun. Reducing Approximation and Estimation Errors for Chinese Lexical Processing with Heterogeneous Annotations[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics. 2012, 232–241.

[13] Collobert, Ronan and Weston, Jason. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]// Proceedings of the 25th international conference on Machine learning. 2008, 160–167.

[14] Huang, Liang and Sagae, Kenji. Dynamic Programming for Linear-Time Incremental Parsing[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics. 2010, 1077–1086.

[15] Zhang, Kaixu and Wang, Ruining and Xue, Ping and Sun, Maosong. Extract Chinese Unknown Words from a Large-scale Corpus Using Morphological and Distributional Evidences[C]// Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. 2011, 837–845.