基于字的分布表征的汉语基本块识别

李国臣[1]，党帅兵[2]，王瑞波[3]，李济洪[3+]

（1.太原工业学院，山西太原 030008；2. 山西大学计算机与信息技术学院，山西太原 030006；3. 山西大学计算中心，山西太原 030006；）

摘要：汉语的基本块识别是汉语句法语义自动分析中的重要任务之一。传统的方法大多数直接将汉语基本块识别任务转化成词层面的一个序列标注问题，采用 CRF 模型来处理。虽然，在许多评测中得到最好的结果，但基于词为标注单位，在实用中受限于自动分词系统以及汉语词特征的稀疏性。为此，本文给出了一种以字为标注单位，以字为原始输入层，来构建汉语的基本块识别的深层神经网络模型，并通过无监督方法，学习到字的 C&W 和 word2vec 两种分布表征，将其作为深层神经网络模型的字的表示层的初始输入参数来强化模型参数的训练。实验结果表明，使用五层神经网络模型，以[-3,3]窗口的字的 word2vec 分布表征，其准确率、召回率和 F 值分别达到 80.74%，73.80%和 77.12%，这比基于字的 CRF 高出约 5%。这表明深层神经网络模型在汉语的基本块识别中是有作用的。

关键词：汉语基本块；分布表征；深层神经网络；序列标注；

中图分类号：TP391　　**文献标识码：**A

Chinese Base-Chunk Identification Based on Distributed Character Representation

LI Guochen[1], DANG Shuaibing[2], WANG Ruibo[3], LI Jihong[3+]

(1. Taiyuan Institute of Technology, Taiyuan, Shanxi 030008;

2. School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006;

3. Computer Center of Shanxi University, Taiyuan, Shanxi 030006)

ABSTRACT: Chinese base-chunk identification is one important task for automatically syntactic and semantic analysis. A widely-used method is to transform it into a word-level sequence labeling problem, and then use CRFs to deal with it. Despite it has achieved the best results in many evaluations, the application in practical situation is limited by accuracy of Chinese word segmentation systems and sparsity of Chinese word features. Therefore, this paper presents a base-chunk identification model based on deep neural network models, which take Chinese character as tagging unit and original input layer. Moreover, Chinese character's C&W distributed representation and word2vec distributed representation are derived through unsupervised learning models, and they are taken as initial input parameters of deep neural network to improve the training procedure. Experimental results show that the precision, recall and F-value of our final identification model can achieved 80.74%, 73.80% and 77.12% respectively, conditioned on a five-layer neural network with feature window of size [-3, 3] and word2vec distributed representation. the F-value is about 5% higher than character-based CRF model. This illustrates that deep neural network model has significant effect on the task of Chinese base-chunk identification.

KEY WORD: Chinese base-chunk , distributed representation ,deep neural network, sequence labeling.

1. 引言

汉语句法分析体系，目前主要有两种：第一种是直接在分词、词性标注的基础上构建汉语句子的完全句法分析树，另外一种是将汉语句子分割成不同层面的语块的浅层句法分析。后者的典型代表是周强等提出的语块分析体系[1]。该体系中提出了汉语基本块、多词块和功能块三种块。周强构建了相应语料，并设置了汉语基本块等自动识别任务，开发了自动分析

收稿日期：　　**定稿日期：**

基金项目：国家自然科学基金（60873128）；山西省科技基础条件平台建设项目（2013091003-0101）

工具。周强构建的第一个基于规则的汉语基本块分析器[2]，在其测试集上 F 值达到 89.47%。不过，该基本块分析器十分依赖于汉语句子的分词和词性标注性能。后来，宇航等[3]使用条件随机场模型构建了一个汉语基本块标注模型，模型的 F 值达到 89.54%。在周强组织的汉语基本块分析评测 CIPS-Pars-Eval-2009 中，基本块自动识别的最好结果为 F 值[4]93.20%（封闭测试）和 90.53%（开放测试）。

需要注意的是，上述评测任务中，绝大多数是通过使用最大熵、条件随机场等模型对汉语句子中的每个词语进行标注，直接将词语、词性等原子特征及组合特征加入到学习算法中进行训练，并最终将预测得到的标记合并成汉语基本块的识别结果。这种做法存在两个问题：第一、这些基本块识别模型的性能非常依赖于测试集中分词的正确性和一致性。设想如果测试集中使用的分词规范和训练集不一致时，基本块自动识别的性能会有着很大的下降。第二、这些基本块识别模型主要使用词、词性等示性特征，机器学习算法很难学习、泛化相同或相近句法结构而使用不同词语表达的样例。

第一个问题的一种解决思路是避免使用词语作为标注单位，而直接使用汉字作为标注单位。这样可以避免由于分词错误或者不一致而导致的基本块标注的性能明显下降。目前，已经有很多的研究工作在探究直接从汉字出发来识别句法块，构建汉语句法分析树[5]。本文采用这种方法，直接将汉语基本块看作是以汉字为标注单位的任务，并使用条件随机场、最大熵和深层神经网络等机器模型来进行标注。

在第二个问题中，我们可以使用几种方法来将词语之间的句法、语义关联信息加入到机器学习算法中。其中，一种方法是，直接使用知网、同义词词林等人工构建的语义资源，以这些资源构建特征加入到机器学习算法中，来提高模型识别的性能[6]。另外一种方法是，使用潜在语义分析，PLSA[7]以及 LDA[8]等算法在使用大规模生语料库训练出各个词语的实值向量表示，并将这些表示作为特征加入到机器学习算法中，来改进模型识别的精度。在本文提出的方法中，我们直接使用两种经典的神经网络模型(C&W[9]和 word2vec[10][11][12])在大规模语料上进行无监督的训练，得到汉字的分布表征，并将这种表征加入到基于字的汉语基本块识别模型中，来验证该分布表征信息对模型性能的影响。

使用神经网络来获取字以及词的分布表征信息已经得到了研究者的广泛关注。其中，最著名的是 Bengio 等人[13]提出的语言模型。该工作中，将英文句子中词语的 n-gram 串通过一个实值矩阵映射成一个固定维度的实值向量，然后将其作为输入，使用神经网络模型构建了一个概率语言模型。在大规模语料上进行训练中，不断地对实值矩阵中的各个元素进行更新学习，最终形成了每个词语的分布表征。在 Collobert 和 Weston 等人[9]的工作中，通过替换 n-gram 词串的当前词来构造出一些伪例，然后将真实的 n-gram 串和构造的伪例作为训练样本，使用 hinge 损失函数来无监督地训练整个神经网络模型，获得了英文词语的分布表征。后来的很多研究工作将这种方法获得的分布表征称为 C&W 分布表征。另外一个著名的工作是 Mikolov 等人提出的[10]。该工作中提出的 CBOW 方法和 Skip-gram 算法具有训练速度快、分布表征性质良好等特点。这些工作中有很多的例子表明，使用大规模无监督的语料进行训练后，词语的分布表征可以较好地体现原词语的句法、语义信息的相似性。本文主要使用了汉语字的 C&W[9]分布表征和 word2vec[10][11][12] 分布表征（使用 CBOW 方法获得）。

字和词的分布表征也被很多研究者使用到自然语言处理的各种任务之中，例如，英文的情感分析、词性标注、命名体识别、语义角色标注以及汉语的分词[14]、基本块识别[15]等任务中。Collobert 和 Weston 等人的研究工作[9]将英文中的词性标注、名命体识别和语义角色标注等多个任务直接放入到一个神经网络模型中，使用分布表征矩阵来将英文词映射到实值向量上，并使用梯度下降算法进行训练，得到了一个接近于目前最好性能的自然语言理解模型。Turian 在文献[16]中提出一种适用于自然语言理解任务的半监督学习框架，即：将无监督训练得到的词语的分布表征作为特征加入到有监督的机器学习算法中，来改进各种自然语

言理解模型的性能。来斯惟等人使用字的分布表征和神经网络算法来构建汉语分词模型[14]。他们的实验结果表明，该方法在汉语分词任务上有很大的潜力。侯潇琪等人[15]将词的分布表征加入到基本块识别模型中，在正确分词基础上 BIO 的标记精度达到 85.90%的。不过，该工作使用词作为标注单位，实用中标注结果明显依赖于分词性能的好坏。

本文直接将字作为标注单位来构建基本块识别模型。在仅仅使用以字构建的特征下，本文对比了条件随机场、最大熵和深层神经网络等标注模型，并对比了字的随机向量表示、C&W 表示和 word2vec 表示三种分布表征。实验结果表明，在[-3,3]窗口下，将字的 word2vec 分布表征融入到五层神经网络下，汉语基本块的识别性能最好，可以达到 77.12%的 F 值。本文的主要目的是基于汉语基本块识别任务，探讨汉语词语的表示学习以及深层神经网络语言模型的有效性。

本文章节安排如下：第 2 节介绍了本文的整个基本块识别模型框架，并详细给出了本文使用的深层神经网络的具体配置以及标注算法所使用的标记集合；第 3 节描述了本文所用的实验数据、实验设置和评价指标；第 4 节总结了实验结果，并进行了深入的分析；最后对本文工作做了总结，并给出下一步的研究方向。

2. 基于字的汉语基本块识别模型描述

本文将基本块识别转化成汉字的序列标注任务，然后借助于多种统计机器学习算法对该序列标注问题进行建模。

2.1 问题描述

汉语基本块识别任务是对给定的一个汉语句子，标注每个基本块的位置，确定基本块中所包含的具体词语。由于一个句子中的汉语基本块不存在重叠、嵌套和交叉问题。因此，我们可以很容易地将其转化成一个序列分割问题，数学描述如下：

将一个汉语句子 $X=(x_1x_2\ldots x_n)$ 看作是由字组成的一个 n 长的序列，每个字使用 x_i 表示。将句子 X 中包含的 p 个基本块看作是一个分割集 $S=\{s_1s_2\ldots s_p\}$。其中，每一个分割 $s_j=<t_j,u_j>$，即，第 s_j 个分割是从第 x_{tj} 个字开始，到第 x_{uj} 个字结束，并且 $u_j\geq t_j$。对于任意两个分割 s_j 和 s'_j，需要满足 $x_j>u'_j$ 或者 $x'_j>u_j$。基本块识别的任务就是，给定 $X=(x_1x_2\ldots x_n)$，正确的识别出分割集 $S=\{s_1s_2\ldots s_p\}$。

上述的序列分割问题，通过要引入一个标记集合来将一个分割的识别问题转化到分割中所包含字的标注问题，即：给分割中包含的每一个字赋予一个标记来标识该字在分割中的位置。常用的标记集合有 IOB1, IOB2, IOE1,IOE2, IOBES 等，具体的转化方法请参见文献[17]。本文中采用了 IOBES 标记集合。其中用“S”标记单字基本块，对于包含多个字的基本块，块中的第一个字用“B”标记，最后一个字用“E”标记，中间的字用“I”标记，对于块外的字统一用“O”标记。具体的对应关系如下例所示：

原始句子:　　医和药是密切相关的。

基本块信息：[医] 和 [药] [是] [密切相关] 的 。

标记信息:　　医/S 和/O 药/S 是/S 密/B 切/I 相/I 关/E 的/O 。/O

通过转化，基本块识别问题可以转化成一个序列标注问题：给定汉语句子序列 $X=(x_1x_2\ldots x_n)$，正确识别出一个句子的基本块信息标记序列 $Y=y_1y_2\ldots y_n$，其中，y_i 属于{I,O,B,E,S}。即找到:

$$Y^*=\underset{Y}{argmax}P(Y|X=x_1x_2\ldots x_n) \quad (1)$$

$s.t.$　Y^* 是一个合理的序列，可以还原出基本块信息

2.2 深层神经网络模型

解决(1)式所描述的问题，条件随机场模型[18]是一种较好的算法。不过，本文仅在一部分对比实验中使用了条件随机场模型。本文主要关注最大熵模型和深层神经网络模型。这两种模型均把(1)式描述的原始问题转化成如下的问题：

$$Y^* = y_1^* y_2^* \dots y_n^* \tag{2}$$

$$y_i^* = \underset{y_i}{argmax} P(y_i | X = x_1 x_2 \dots x_n) \tag{3}$$

$s.t.$ Y^*是一个合理的序列，可以还原出基本块信息

实际上，只有当标记序列 Y 中任意两个 y_i 和 y_j (i 不等于 j)之间相互独立时，(1)式才可以转化成(2)和(3)式。在本文中，为了简单处理，我们假设这种独立性成立。最大熵算法的基本思想和模型形式在文献[19]中已经给出了很好地描述。这里，我们仅给出本文使用的深层神经网络模型的结构及一些参数设置。

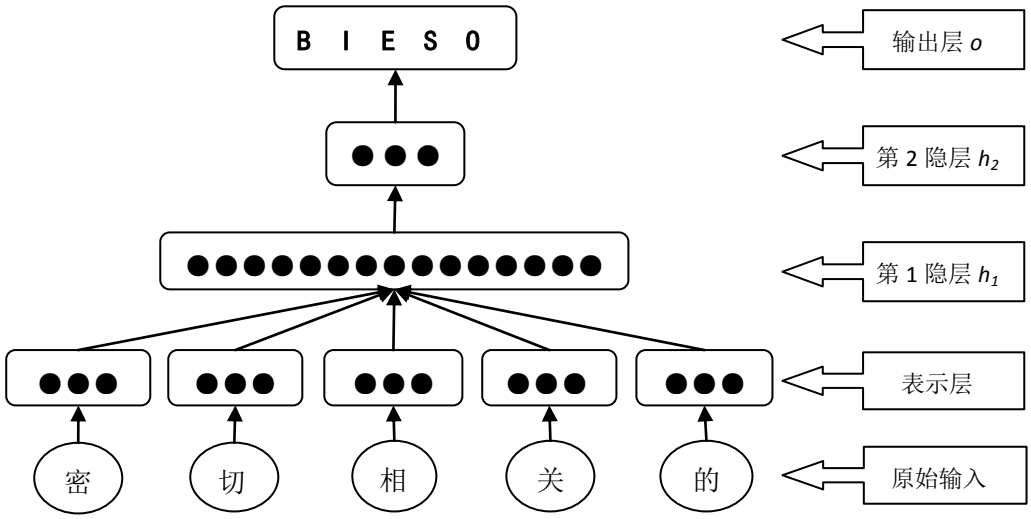

图 1. 深层神经网络模型图

本文所使用的深层神经网络为五层结构(算上原始输入层)，如图 1 所示。其中，原始输入是 w 个字在字表中对应的索引值，然后通过在分布表征矩阵中查找出这 w 个字中每个字所对应的 n 维的实值向量，并依字序首尾相接成的 $w \cdot n$ 维的实值向量 v。在第 1 隐层直接使用 tanh 函数对向量 v 进行非线性变换形成 h_1，该层中的每个节点 h_{1j} 都与表示层中的每个节点 v_i 相连。进而，在第 2 隐层中，将 h_1 使用 tanh 变换得到 h_2。同样，h_1 和 h_2 之间的节点也是完全相连的。最后，在输出层，使用 softmax 函数对 h_2 层的节点值进行概率归一化得出 $P(y_i = t \mid X)$，t 属于{B,I,O,E,S}。最终，使用输出层的 5 个节点中最大概率值对应的标记作为第 i 词的最终标记。

2. 3 字的分布表征

一般来说，在使用机器学习算法解决(1)和(3)中描述的问题时，并没有将整个句子 X 中的所有信息引入来预测每个字的标记信息。在预测第 i 个字的标记时，通常只是将该字周围的一些字的信息作为特征加入到机器学习算法中，即，使用开窗口的方式来进行特征选取。本文也采用了这种做法。

区别于直接使用字、词作为特征，本文使用了字的分布表征作为特征信息。不同于字的 0-1 向量表示，字的分布表征是将字表示成为一个定长的实值向量。该实值向量是通过某个表示学习模型来获得的。具体请参考本文第一部分给出的相关文献。

需要注意的是，假设常用字有 5000 个，并且特征窗口设置为[-2,2]，如果直接使用字的 0-1 向量表示作为特征，那么机器学习算法就必须处理 25000 维的特征向量。如果再考虑上字、词特征之间的组合特征等，那么特征向量的维度会成倍的往上增长。这很容易引起维数

灾难问题。并且，在这样的特征矩阵里，存在着严重的稀疏问题。字的分布表征一般为低维（比如 100 维）的实值向量，那么上面的问题就可以转化为一个在 500 维特征上的一个学习问题。值得说明的是，字的分布表征中能学习到字之间的句法、语义的某些关联性，这为自然语言处理许多任务带来新思路、新方法。下面的几节中，给出汉语基本块识别任务实验。

3. 实验设置

本文实验主要关心使用字的分布表征来进行基本块的识别实验。在所有实验中，基本块的识别问题被转化成字层面的序列标注问题。本文主要使用了 IOBES 标注集合。实验中尝试了多种机器学习算法，并对它们进行了比较。

3.1 实验语料

实验语料使用了 CIPS-ParsEval-2009 中发布的汉语基本块分析语料。语料库总规模为 765820 字 ，训练文件数为 171 个, 包含 14249 条句子, 共计 618231 字。测试文件数为 43 个, 包含 3751 条句子 ,共计 147589 字。语料中，基本块的块长（块中所含字数）统计如表 1 所示。

表 1: 基本块语料块长度统计

块长	1	2	3	4	5	6	7	>7
块数	38053	104595	28171	39899	8072	7680	2220	2054
所占比例	16.49%	45.33%	12.21%	17.29%	3.50%	3.33%	0.96%	0.89%

从表 1 中可以看出，块长小于等于 5 的块占到近 95%，而长度不大于 7 的块占所有块的 99%。因此，在实验中，本文分别使用[-2,2]和[-3,3]窗口内的字的分布表征作为特征，来对当前字进行标注。

3.2 要对比的标注模型

本文使用了三种标注模型：最大熵、条件随机场和深层神经网络。其中，条件随机场模型在序列分割和标注任务中得到了广泛的应用[18]。本文使用了张乐博士开发的 MaxEnt 最大熵工具包[20]。在将字的分布表征作为特征值代入到最大熵工具时，做了平移处理（加上某个常量）让所有的值都转变为正数。实验中，高斯惩罚参数设为 1.0。本文使用的深层神经网络模型实在 pylearn2 工具包[21]上开发得到的。本文主要构建了一个五层神经网络模型，该模型的结构在 3.2 节中给出。其中，第一个隐层的单元个数为 300，第二个隐层的单元个数为 100。另外，本文进行对照实验使用条件随机场模型的 crfpp 工具包[22]。

3.3 字的分布表征学习算法

本文使用C&W算法[9]和 word2vec工具包[10][11][12] 的CBOW算法来获得字的分布表征。其中，我们设置每一个字使用 100 维的实值向量来表示。两种工具包的训练语料均使用的是山西大学 500 万分词语料。学习分布表征前，我们对语料库进行了简单的预处理，把所有的英文字母统一用“WORD”表示，所有的数字用“NUMBER”表示。在进行基本块识别前，我们将每一个字的分布表征单位化成一个长度为 1 的向量。

在 C&W 算法中，本文仅将隐层设置为一层，学习率设置为 0.00000001，迭代时使用的是句子中字的 5 元组作为原始输入。模型使用 BGD(Batch Gradient Descent)优化算法，其中，每一个 minibatch 设置为 1000。由于该表示学习算法可以无限地迭代下去，本文仅选择迭代到 5500 万 minibatch 后生成的字的分布表征。

在 word2vec 工具包中，本文使用的是 CBOW 算法，并且使用层次化的 softmax 层作为输出层，在训练时设置窗口大小为 5。

为了观察 C&W 方法和 word2vec 方法的训练效果，本文仿照文献[14]，选取了“一”，“李”，“江”，“急”四个字，并给出了它们的最相似字。这里，我们先将字的分布表征向量

进行单位化，然后使用夹角余弦计算相似度。具体结果见表 2。

表 2：不同字表示学习方法得到的“一”“李”“江”“急”的最相似字

C&WC				Word2Vec			
一	李	江	急	一	李	江	急
这	王	北	惨	这	鹏	浙	紧
三	刘	南	死	两	刘	湖	救
两	吴	西	哀	都	秉	省	迫
那	陈	湖	错	就	邹	岷	忙
几	赵	燕	忧	几	赵	河	住
大	沈	毛	尽	了	俊	澧	诊
多	朱	河	遭	那	孙	陕	抓
各	秦	黄	无	，	玲	川	痛

从表 2 中可以看出，C&W 和 word2vec 两种方法学习到的字的分布表征还是有所差别的。从直觉来看，C&W 方法对“李”的聚类结果要比 word2vec 方法的要好。而对于“一”，“江”和“急”，两种方法的聚类结果尽管不尽相同，但是，并没有明显的好坏之分。

3.4 评价指标

本文从字层面和块的层面来评价基本块识别模型的性能。其中，在字层面，本文使用了**标记准确率，**它指的是所有标签中标记正确的标记数与总的标记数的比值。在块层面，本文使用了块识别的准确率、召回率和 F 值。它们的定义如下：

准确率 = 识别出的正确块数/识别出的总块数

召回率 = 识别出的正确块数/测试集中的总块数

F 值 = 准确率*召回率* 2 /（准确率+召回率）

4. 实验结果和分析

本节中，我们首先分析了不使用字的分布表征，只使用字作为特征的各种基本块识别模型性能，然后又分析了使用字的分布表征的各模型性能，最后对基于词的神经网络模型与基于字的神经网络模型做了对比分析。

4.1 不使用字的分布表征的结果

我们直接将字作为特征代入到基本块识别模型中。表 3 和表 4 分别给出了两种学习算法使用字特征时的基本块识别性能。

表 3: MaxEnt 算法+字特征

窗口大小	标记准确率	准确率	召回率	F 值
[-2,2]	83.96%	73.48%	64.37%	68.63%
[-3,3]	84.11%	73.54%	64.86%	68.93%

表 4: Crfpp 算法+字特征

窗口大小	标记准确率	准确率	召回率	F 值
[-2,2]	85.95%	73.09%	71.31%	72.19%
[-3,3]	85.96%	73.03%	71.54%	72.28%

对比表 3 和表 4 可以发现，两者的实验结果都较目前较好的一些基本块分析模型的性能[4]要差很多。这主要是因为上述实验中并没有考虑词性特征和词、词性的组合和搭配特征，而这些特征的加入可以明显改善基本块识别的性能。之所以不加入这些特征，主要是本文旨在探讨字的分布表征对基本块识别的影响。

表 5 中给出了将词作为标注单位，将[-2,2]窗口内的词特征加入到条件随机场模型中，进行基本块识别的结果。为了对比字特征与词特征对于基本块识别性能的影响，我们在实验中也未使用词性特征，以及多元的组合搭配特征。

表 5: 词为标注单位+Crfpp+[-2,2]窗口

分词方式	准确率	召回率	F 值
正确分词	81.89%	82.47%	82.18%
自动分词	65.67%	68.24%	66.93%

从表 5 中可以看出，如果分词信息正确，基本词层面的块识别 F 值可以达到 82.18%。但是，当使用山西大学分词软件 FC2000 对测试集的句子自动分词后，基本块识别的 F 值仅可以达到 66.93%，明显低于表 4 中给出的以字为标注单位的实验结果。这说明以词语为标注单位的基本块识别模型在实际使用中，对于分词系统的性能有着很大的依赖性。这也是本文希望研究以字为标注单位的基本块识别模型的重要原因之一。

4. 2 使用字的分布表征的结果

这一小节，我们将字的分布表征分别加入到最大熵模型，CRF 模型和深层神经网络模型后的实验结果。

4.2.1 最大熵+字的分布表征

为了对比验证，我们将如下三种字的分布表征加入到最大熵模型中。三种分布表征中，除了包含上文提到的 C&W 字表示和 word2vec 字表示，本文还加入了完全随机的字表示。随机字表示是针对每一个字随机生成了一个 100 维的实值向量。向量中的每一个元素从[-0.01,0.01]的均匀分布中抽取，然后，对该向量进行单位化。

表 6 给出了将字的三种分布表征加入到最大熵算法中的基本块识别结果。

表 6: MaxEnt 算法+字的分布表征

窗口	字分布表征类型	标记准确率	准确率	召回率	F 值
[-2,2]	随机字表示	59.29%	45.13%	28.63%	35.04%
	C&W	74.80%	57.31%	44.61%	50.17%
	word2vec	69.97%	52.51%	41.11%	46.12%
[-3,3]	随机字表示	59.09%	42.38%	27.45%	33.31%
	C&W	75.55%	59.11%	47.48%	52.66%
	word2vec	70.19%	54.43%	39.62%	45.82%

对比表 6 中的三种分布表征的实验结果，可以发现，虽然 C&W 表示特征和 word2vec 表示特征的识别结果较完全随机的表示特征有着明显的上升(F 值上升近 10%~15%)，但识别结果也很不理想（F 值仅在 50%左右）。探究其原因，从分类算法的角度来看，主要因为最大熵分类器并不考虑整个序列的全局优化，仅是针对每个字的标记的单点优化；从特征的表示来看，C&W 和 word2vec 的分布表征尽管克服了原有的 0-1 表示特征的数据稀疏问题，但是，两种分布表征是使用无监督的方式训练得到的，而没有针对具体任务进行优化，因此，它们并没有很好地表达出基本块识别所需要的句法语义信息。

对比表 6 中的两种窗口下的实验结果，可以发现，窗口的扩大并没有带来识别结果的明显提升，甚至在随机分布表征和 word2vec 分布表征的来中情况下，窗口的扩大还带来了块 F 值的些许下降。

4.2.2 CRF+字的分布表征

为了与基于字特征的 CRF 模型作对比，我们把上述三种分布表征作为特征直接应用到

CRF 模型中。表 7 是得到的详细结果。

表 7: CRF 算法+字的分布表征

窗口	字分布表征类型	标记准确率	准确率	召回率	F 值
[-2,2]	随机字表示	65.04%	43.63%	40.03%	41.75%
	C&W	76.44%	57.66%	54.72%	56.15%
	word2vec	70.77%	49.74%	46.03%	47.81%
[-3,3]	随机字表示	65.14%	44.11%	40.73%	42.35%
	C&W	76.14%	57.70%	54.58%	56.10%
	word2vec	70.62%	49.98%	46.27%	48.05%

对比表 7 和表 4，可以发现，使用字的分布表征时，其最好结果也明显低于仅使用字特征的 CRF 模型。

4.2.3 深层神经网络+字的分布表征

这一小节，我们给出了使用深层神经网络来进行基本块识别的实验结果。表 8 中详细总结了在两种窗口下三种分布表征的条件下，基本块识别的详细结果。

表 8: 深层神经网络+字的分布表征

窗口	字分布表征	标记准确率	准确率	召回率	F 值
[-2,2]	随机字表示	87.61%	78.90%	71.62%	75.09%
	C&W	87.74%	79.90%	71.25%	75.33%
	Word2Vec	87.63%	79.71%	71.23%	75.23%
[-3,3]	随机字表示	88.34%	78.87%	74.33%	76.53%
	C&W	88.37%	80.37%	73.04%	76.53%
	Word2Vec	88.67%	80.74%	73.80%	77.12%

分别对比表 8 和表 6、表 7，可以看出，五层的神经网络模型的实验结果明显好于最大熵模型和 CRF 模型。这一方面得益于深层神经网络模型使用多个非线性隐层来对原始的分布表征进行变换，形成更为有用的特征。另一方面体现出深层神经网络在进行基本块的学习过程中，对原有分布表征进行调整，形成了对基本块识别任务更为有利的分布表征。

对比表 8 中三种分布表征的影响，可以看出，尽管基于 C&W 分布表征和 word2vec 分布表征所得到的基本块识别结果都比使用完全随机的分布表征要好一些，但是三种分布表征的实验结果之间的差异不大，word2vec 分布表征的结果略高一些。也就是说，三种表示作为深层神经网络的初始输入，对最后的基本块识别影响不大。这也说明，神经网络模型在迭代计算的过程，对字的分布表征进行不断地学习、修正，弱化了对初始值的依赖，形成了基本块识别任务需要的字的分布表征。

在表 8 中，不同窗口的实验结果表明，扩大特征窗口可以带来识别结果的明显上升。而且在[-3,3]窗口内，使用 word2vec 分布表征可以达到本文最高的识别 F 值，即 77.12%。这明显好于表 4 中给出的条件随机场情况下使用字特征得到的实验结果(F 值为 72.28%)。需要强调的是，和最大熵模型一样，本文使用的深层神经网络模型在训练也是仅针对每个字的单点标记似然进行最大化，而不是优化整个序列上的似然函数。因此，使用深层神经网络的识别结果能高出条件随机场识别结果近 5%的 F 值也是相当可观的。

另外，本文使用 word2vec 分布表征，分别用 4 层、6 层神经网络也做了实验，其结果均低于 5 层神经网络模型，但差异不大（见表 9）。这说明，在汉语基本块识别任务中选择 5 层神经网络是合适的。从语言层面来分析，可以将 h_1 隐层理解为关于词的特征表示，h_2 隐层可以理解为关于基本块的特征表示。字的分布表征是经过词的特征表示再到基本块的特征

表示，或略掉词的特征表示层（h_1 隐层）直接到基本块的特征表示层（h_2 隐层），即用 4 层神经网络，是不可取的。同样，多于 5 层时模型结构难以从语言层面合理解释，相应的标注结果也有所下降。

表 9: 使用 word2vec 的 4 层和 6 层神经网络的结果

层数	标记准确率	准确率	召回率	F 值
4	88.24%	80.10%	72.96%	76.36%
6	88.45%	80.48%	73.42%	76.79%

4. 3 基于词的深层神经网络模型结果

表 10 给出了基于词的神经网络模型结果，为了与基于字的神经网络模型作对比，测试集分别使用了原人工标注的正确分词语料和经过山西大学分词软件 FC2000 重新分词后的语料。

表 10: 基于词的神经网络模型

窗口大小	分词方式	准确率	召回率	F 值
[-2,2]	正确分词	85.02%	81.97%	83.47%
	自动分词	73.15%	75.71%	74.41%
[-3,3]	正确分词	83.79%	82.07%	82.92%
	自动分词	73.31%	75.01%	74.15%

对比表 10 和表 5 可以看出，本文所用到的神经网络模型性能要优于 CRF 模型，这也与上一小节得到的结论一致。比较表 10 和表 8 可以看出当测试集使用自动分词语料时其结果要低于基于字的神经网络模型，这也在神经网络模型上验证了 4.1 小节由表 5 得到的结论。

5. 总结与展望

本文研究了和对比了使用字的分布表征来进行基本块识别的若干种方法。在这些方法中，本文主要使用了最大熵、条件随机场和深层神经网络三种模型，并且使用了字的 C&W 分布表征、word2vec 分布表征、随机的字分布表征，在[-2,2]和[-3,3]两种特征窗口情形下，我们对多个基本块识别模型进行了对比。实验结果表明，使用在[-3,3]窗口下，将字的 word2vec 分布表征融入到五层神经网络模型下，可以得到汉语基本块的一个较好的识别性能（F 值达到了 77.12%）。这个结果要明显好于直接将[-3,3]窗口内的字特征加入到条件随机场模型所得到的识别模型(F 值为 72.28%)。

实际上，本文的所有实验中并未能融入词性信息、字的组合搭配信息等更为丰富的特征信息。我们相信如果将这些信息进一步加入到本文的模型中，基本块的识别性能还会有大幅度的提高。但如何获得词性的分布表征以及相邻字的组合串的分布表征是需要我们进一步研究的。

参考文献

[1] 周强，任海波，孙茂松. 分阶段构建汉语树库[A]. In Proc. of The Second China-Japan Natural Language Processing Joint Research Promotion Conference, 2002:189-197.

[2] 周强. 基于规则的汉语基本块自动分析器[C]. 第七届中文信息处理国际会议论文集(ICCC-2007). 2007: 137-142.

[3] 宇航，周强. 汉语基本块标注系统的内部关系分析[J]. 清华大学学报，2009, 49(10):136- 140.

[4] 李超，孙健，关毅，徐兴军，侯磊，李生. 基于最大熵模型的汉语基本块分析技术研究[R]. CIPS-ParsEval -2009,2009.

[5] 赵海，揭春雨，宋彦. 基于字依存树的中文词法-句法一体化分析[C]. 全国第十届计算语言学学术会议(C- NCCL-2009), 2009:82-88.

[6] 齐璇,王挺,陈火旺. 义类自动标注方法的研究[J]. 中文信息学报,2001,15(3):9-15.

[7] 吴志媛，钱雪忠 .基于 PLSI 的标签聚类研究[J]. 计算机应用研究，2013,30(5).

[8] David M. Blei. Latent Dirichlet Allocation[J].Journal of Machine Learning Research,2003(3):993-1022.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research (JMLR), 2011:2493-2537.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space [J]. arX- iv preprint arXiv,2013:1301-3781.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed representations of words and phrases and their compositi- onality [J]. arXiv preprint arXiv,2013:1310-4546.

[12] Tomas Mikolov,Wen-tau Yih, and Geoffrey Zweig.Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.

[13] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. Journal of Machine Learning Research (JMLR),2003:1137–1155.

[14] 来斯惟,徐立恒,陈玉博,刘康,赵军. 基于表示学习的中文分词算法探索[J]. 中文信息学报,2013,5(9):8-14.

[15] 侯潇琪，王瑞波，李济洪. 基于词的分布式实值表示的汉语基本块识别[J]. 中北大学学报（自然科学版）. 2013,34(5):582-585.

[16] Turian Joseph, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). 2010.

[17] Taku Kudo and Yuji Matsumoto. Chunking with support vector machine. In Proceedings of the second meeting of North American chapter of association for computational linguistics(NAACL), 2001: 192-199.

[18] John Lafferty, Andrew Mccallum , FernandoPereira. Conditional random fields :Probabilistic models for segmenting and la- beling sequence data[C].International Conferenceon Machine Learning (ICML 01) . William stown , MA ,U- SA , 2001 : 282-289.

[19] Berger Adam, Stephen Della, Pietra Adam, Vincent Della Pietra. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1):39-71.

[20] 张乐. 最大熵工具包 MaxEnt(2004 版)[CP/OL].2004.http://homepages. inf.ed.ac.uk/s0450736/maxent_ toolkit .html.

[21] Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. “Pylearn2: a machine learning research library”. arXiv preprint arXiv:1308.4214.

[22] TakuKudo, CRF++ toolkit, 2005. http://crfpp.sourceforge.net/.