基于表示学习的中文分词算法探索

来斯惟，徐立恒，陈玉博，刘康，赵军

中国科学院自动化所模式识别国家重点实验室，北京，100190

E-mail: {swlai, lhxu, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

摘　要：分词是中文自然语言处理中的一个关键基础技术。通过基于字的统计机器学习方法学习判断词边界是当前中文分词的主流做法。然而，传统机器学习方法严重依赖人工设计的特征，而验证特征的有效性需要不断的尝试和修改，是一项费时费力的工作。随着基于神经网络的表示学习方法的兴起，使得自动学习特征成为可能。本文探索了一种基于表示学习的中文分词方法。首先从大规模语料中无监督地学习中文字的语义向量，然后将字的语义向量应用于基于神经网络的有监督中文分词。实验表明，表示学习算法是一种有效的中文分词方法，但是我们仍然发现，由于语料规模等的限制，表示学习方法尚不能完全取代传统基于人工设计特征的有监督机器学习方法。

关键词：表示学习、中文分词

Chinese Word Segment Based on Character Representation Learning

LAI Siwei, XU Liheng, CHEN Yubo, LIU Kang, ZHAO Jun

National Laboratory of Pattern Recognition Institute of Automation, Beijing 100190

E-mail: {swlai, lhxu, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract: Word segmentation is a fundamental technology of Chinese natural language processing. Using character-based statistical machine learning methods to perform Chinese word segmentation is the main trend currently. However, conventional machine learning methods heavily rely on manually designed features, which require intensive labor to modify the features and verify their effectiveness. With the rapid develop of neural-network-based representation learning, it becomes realistic to learn features automatically. This paper investigates a Chinese word segment method based on representation learning. We first learn embedding vectors for Chinese characters from a large corpus unsupervisedly, and then apply them to neural-network-based Chinese word segmentation supervisedly. Experimental results show that representation learning is an effective method for Chinese word segmentation. However, due to the limitation of corpus size, it still cannot replace conventional machine learning methods which based on manually designed features.

Keywords: Representation Learning, Chinese Word Segmentation

1　引言

词是“最小的能独立运用的语言单位”[1]，由于中文具有大字符集连续书写的特点，如果不进行分析，计算机则无法得知中文词的确切边界，从而很难理解文本中所包含的语义信息。因此，中文分词是自然语言处理中的一个关键的基础技术，是其他中文应用，例如命名实体识别、句法分析、语义分析等，的前期文本处理关键环节，其性能的优劣对于中文信息处理尤为重要。

传统对于中文分词的研究比较丰富，例如：最大正向匹配、最大逆向匹配、双向匹配等基于词典的匹配方法。然而，由于语言的复杂性，中文文本中存在大量的词边界歧义与未登录词(OOV)。仅仅是基于词典的匹配方法无法有效地解决以上两个中文分词中的关键难点问题。所以越来越多的方法关注基于字的中文分词。基于字的中文分词方法基本假设是一个词语内部文本高内聚，而词语边界与外部文字低耦合。每一个词都可以通过其所在的上下文特

征进行表示，通过统计模型可以很好的判别当前字在构词过程中的作用（词的开始、中间、结束或是单字词）。通过大量实验表明这种基于字的中文分词方法要明显优于基于词典匹配的分词方法。然而，基于字标注的分词方法的问题在于：传统的字表示特征，无论是一元特征（Unigram）或是二元特征（Bigram），都很难有效表示目标字，使得统计模型不能有效地理解每个字的含义。另外，所有的特征表示都是基于词袋子模型，然而这样表示模型有两个较为明显的缺点：1）语义鸿沟问题：通过词袋子模型，我们没法直接知道“麦克风”和“话筒”描述的是同样的事物。2）低频词的问题。在使用词袋子特征训练模型时，低频词由于出现次数较少，往往只被训练的极少的次数，容易造成训练不足，也非常有可能过拟合。因此如何对于中文文本中每个字进行建模，并自动的抽取字的表示特征是基于字表示的分词方法中的一个难点问题。

然而，近些年随着深度学习（Deep Learning）的兴起，特征表示学习（Feature Representation Learning）逐步成为机器学习的一个新兴分支。深度学习是利用深层神经网络自动学习出数据的一种表示。自 2006 年 Hinton[2]提出深度学习后，该方法在语音、图像领域均取得了惊人的成果。已有工作表明，随着网络层数的加深，深度学习算法可以学习出越来越抽象的数据表示。在这种特征的基础上进一步地进行模型的学习，可以显著地提高分类的性能。在自然语言处理任务中，深度学习也已经广泛地应用于命名实体识别（NER）、词性标注（POS Tagging）、情感分类（Sentiment Classification）等任务，并有一定优势。然而在中文分词任务中，还未见针对深度学习的应用研究成果。因此本文试图将深度学习应用与中文分词任务，来探讨其是否可以有效地提高分词的性能。

具体地，我们利用基于词的稠密向量表示方法[3]，将一个字用 n 维实数向量来描述。同时采用 SENNA[4]在海量无标注数据来无监督的训练每个字的稠密特征表示向量，并以此作为特征，应用于分词算法中。经过多组实验比对，我们的方法的效果相对于人工设计特征的最大熵算法有一定的竞争力。

文本章节安排具体如下。第 2 章介绍了分词及词的表示学习的相关工作；第 3 章介绍了基于字表示的分词算法框架；第 4 章介绍一种在大规模语料上无监督学习出字的稠密表示的方法；第 5 章为实验及分析；最后对本文工作进行了总结，并指出将来工作的方向。

2 相关工作

传统分词方法依赖词典匹配，并通过贪心算法截取可能的最大长度词进行有限的歧义消除。常用的贪心策略有正向最大匹配法、逆向最大匹配法和双向匹配等。然而，基于词典方法存在两个明显的缺陷，即不能很好得处理词边界歧义和未登录词（OOV）。为了解决中文分词的这两个关键问题，许多研究工作集中到了基于字标注的机器学习中文分词方法。

基于字的中文分词方法基本假设是一个词语内部文本高内聚，而词语边界与外部文字低耦合。通过统计机器学习方法学习判断词界是当前中文分词的主流做法。现有工作大多使用序列标注模型执行 BMES 标注。Xue 等人提出了基于 HMM 模型的字标注中文分词方法[5]。刘群等提出一种基于层叠隐马模型的汉语词法分析方法[6]。该方法引入角色 HMM 识别未登录词，使用 Viterbi 算法标注出全局最优的角色序列。同时，该方法还提出了一种基于 N-最短路径的策略进行切分排歧。Wang 等人使用基于字分类的 CRF 模型进行中文词法分析[7]。对基于字标注中文分词方法的改进包括引入更多的标签和设计更多高效的特征[8][9]、联合使用产生式模型和判别式模型以融合两者的优点[10]以及将无监督方法中使用的特征引入有监督方法中[11]等。然而，传统统计机器学习方法往往依赖于人工设计的特征，而一个特征是否有效需要多次尝试与选择。因此人工设计一系列好的特征既费时又费力。

近年来，随着深度神经网络优化方法的突破[2]，基于神经网络的表示学习方法得到了蓬

勃的发展。在自然语言处理领域，表示学习的目标是要将最小的语义单位表示成一个 n 维向量，向量中的每一维表示某种隐含（latent）的句法或语义信息。Collobert 等人在 2011 年发布了首个基于表示学习的多任务学习系统 SENNA[4]。它将词性标注、命名实体识别、句法分析和语义角色标注任务融合与一个框架，运用神经网络替代传统序列标注模型，进行自动的特征学习，从而避免了繁琐的人工特征设计过程。此后，基于深度神经网络的表示学习方法被应用于句法分析[12]、复述检测[13]、语义分析[14]以及情感分类任务[15]，并取得了巨大的成功：在不需要人工参与设计有效特征的情况下，表示学习方法相比于传统有监督模型取得了等价于或更好的成绩。

3 基于字表示的有监督分词

3.1 数据预处理

中文分词的训练语料中，英文与数字的出现次数较少（甚至有可能 26 个英文字母中有的字母未在训练集中出现过）。为了简化处理流程，本文使用了一个简单的数据预处理步骤，将所有的连续数字字符替换成一个专用的数字标记“NUMBER”，将所有连续的英文字母替换成一个专用的英文单词标记“WORD”。如训练语料“中国/教育/与/科研/计算机网/（/ＣＥＲＮＥＴ/）/已/连接/了/２０００/多/所/大学”经过预处理步骤将会变成“中国/教育/与/科研/计算机网/（/WORD/）/已/连接/了/NUMBER/多/所/大学”。其中 NUMBER 和 WORD 在训练时都当作一个字符来考虑。

这种方法在一定程度上丢失了部分语义信息，会对分词精度产生负面的影响。但是在训练语料不充分的情况下，该预处理可以简化后续步骤，将实验重心放在处理汉字词语上。

3.2 字的稠密向量表示

借鉴 Bengio 等人[3]的思想，本文将每个汉字用一个 *n* 维实数向量来表示（后文简称字向量）。字向量初始化为一个随机的小实数值，在训练过程中，每个字的字向量会进行更新，最后根据训练目标的不同，字向量之间的相似度也会有所不同。具体可见第 4 章实验部分。

3.3 模型及算法

与其它基于字的分词方法相似，本文也采用 BMES 体系对汉字进行标注。对于单字词，其标签为 S；对于多字词，词中的第一个汉字标签为 B，最后一个汉字标签为 E，中间字的标签为 M。对训练数据的每个字进行标注后，本文采用一种 3 层神经网络结构对每个字进行训练，其结构如图 1。

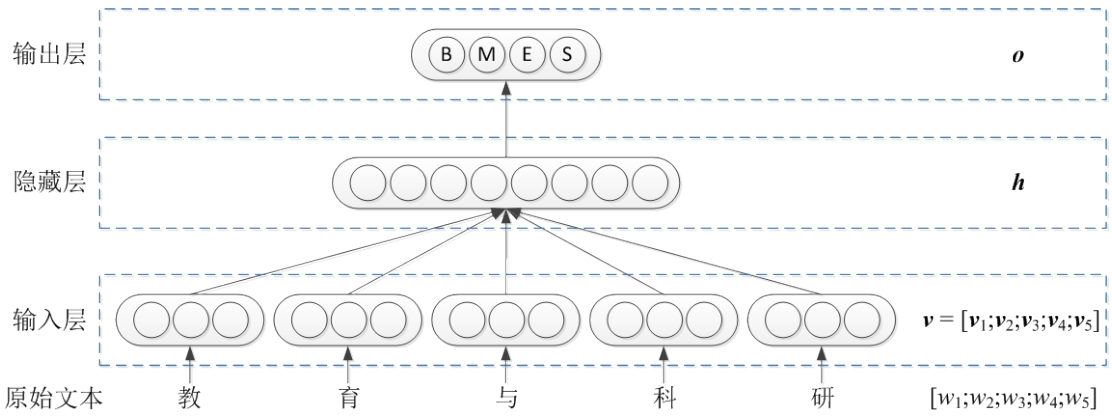

图 1 算法基本结构图

对于句子中的每个字的标签分类任务，本文选取上下文以及当前字，共 w 个字作为特征。其中上文和下文均为 (w-1)/2 个字。图中最下方为这 w 个字的原始文本，经过第一层，将每个字转换成其字向量表示 $\boldsymbol{v}_i$，并把 w 个字连接成一个 wn 维的向量 $\boldsymbol{v}$。该 wn 维的向量是神经网络的输入层。隐藏层 $\boldsymbol{h}$ 的设计与传统的 BP 神经网络一致，输入层的 wn 个节点与隐藏层的 H 个节点之间两两均有边连接。隐藏层选用 tanh 函数作为激活函数。输出层一共有 4 个节点，使用 softmax[16]归一化后，分别表示这个字被打上 B、M、E、S 标签的概率。

网络结构可以形式化的表示为：

$$h = \tanh(\mathbf{U}\boldsymbol{v})$$

$$o = \mathbf{V}\boldsymbol{h}$$

$$p(i|x,\theta) = \frac{e^{o_i}}{\sum_j e^{o_j}}$$

其中 $\mathbf{U}$ 为输入层到隐藏层的权重，$\mathbf{V}$ 为隐藏层到输出层的权重。这两层均可理解为简单的矩阵相乘。最后使用 softmax 函数可以将输出 o 转换成标签概率 p。

网络的训练目标使用传统的最大似然估计法，即求出一组参数θ，并最大化：

$$\sum_{(w,y)\in T} \log p(y|x,\theta)$$

其中参数θ包含各个字的字向量 $\boldsymbol{v}$ 以及两个网络中的参数矩阵 $\mathbf{U}$ 和 $\mathbf{V}$。训练中这里使用了朴素的随机梯度下降法。

4 无监督学习字表示

在有监督的学习中，往往会遇到低频字训练不充分的问题。无论在传统的浅层模型（如最大熵、CRF）中，还是第三章描述的以神经网络为框架的模型中，低频字只会在极少量的样本中出现。因此，如果引入一个更大的语料，从这个语料中学习出各个字更丰富的信息，并加入有监督学习中，将有可能极大地提高有监督学习的训练效果。

4.1 字表示的训练

字表示的无监督训练 Collobert 等人[4]和 Mnih 等人[17]均提出过。Joseph[18]对这两种方法进行了更公平的比较，结果表明，Collobert 的方法略胜一筹。在中文中，尚未看到类似的比较，因此本文直接使用 Collobert 的方法训练字向量。

在无监督字表示训练中，我们仍然使用如图 1 所示的神经网络结构图。不同之处在于，最后一层只输出一个得分，而并不输出 4 个标签的概率。该得分的含义为这个连续的字序列是否是一个正常的词序列。所谓的正常序列，是相对随机序列而言的。语料中真实存在的序列均认为是正常序列，而一个随机从字典中选取若干个字生成的序列，则认为是非正常序列。

由于这个想法与语言模型非常相似，因此在[4]中也被称作语言模型。实际上，Collobert 的方法与传统的语言模型略有差别。传统意义上，语言模型是给定了前若干个字/词，预测下一个字/词。而在这种方法中，并不需要预测下一个字，只需知道一个序列是否是正常序列。在无监督训练阶段，我们希望一个正常的序列可以得到高分，而一个非正常的序列，我们希望它的分数更小。

在实际操作中，正样本可以直接从语料中选取得到，而负样本则需要构造。如果负样本直接从字典中选取若干个随机字符，则容易生成完全没有可读性的字符串。这些字符串会离分类面非常的远，这会造成一些略有“语病”的句子，被分类成正常的句子。为了解决这个问题，使得负样本更接近真实的分界面，本文的负样本由一个真实的序列随机替换一个字得

到。类似的方法在 Collobert 等人[4]和 Turian 等人[18]中提出过。[4]替换了一个序列中最中间的词，而[18]替换了一个序列中最后的词，取得的效果类似。本文在实验中替换的是中间字。

记 x 为一组正常的字序列，则$f_\theta(x)$表示网络的输出。每个负样本记作 x^w，表示一个正常的序列 x 中将中间的字替换为 w。同样的，负样本的输出为$f_\theta(x^w)$。

无监督训练阶段，这里使用成对训练的方法，即最小化如下目标：

$$\sum_{x\in X}\sum_{w\in D}\max\{0,1-f_\theta(x)+f_\theta(x^w)\}$$

式中，X 为从语料集中选取出了所有连续的 w 个字，D 表示字典。

与监督学习阶段相同，这里也采用随机梯度下降法进行训练，最后只使用其词向量部分。

4.2 字表示的使用

通过无监督训练得到的字表示通常有两种用法。一、作为神经网络模型的初始值。二、加入到现有的浅层模型中，如最大熵模型。

[4]将无监督学习得到的词向量作为有监督学习网络中的初始值，大幅度提高了其有监督学习的训练效果。这一思想与音频、图像领域在深度学习中，对深层神经网络的初始值使用受限玻尔兹曼机进行无监督的初始化非常类似。由于神经网络是一个非凸优化的问题，局部极值点非常的多，好的初始值可以使其最后收敛到一个更好的解，同时也能在一定程度上抑制训练的过拟合。

本章无监督训练得到的字向量同样可以直接作为第 3 章中字向量的初始值用于训练。对于网络结构中的 U、V 矩阵，仍然使用随机的初始值。

[18]在英语中使用词向量作为扩展特征，提升了命名实体识别（NER）和语块分析（Chunking）的效果。其方法较为直接，在最大熵做序列标注问题时，直接将周围共 w 个词的词向量直接加入改词特征向量中。

在第 5 章的实验中，我们同时尝试了以上两种思路。

5 实验结果及分析

实验中，我们以最大熵模型作为基准，尝试了本文描述的若干种方法，并进行比较。

在实验中，我们需要确定 w 的大小，即认为上下文窗口中共 w 个字会对当前字的标签产生主要影响。[19]中通过大量实验表明窗口 5 个字可以覆盖真实文本中 99%以上的情况。因此本文也取 w 为 5，即使用上文 2 个字、下文 2 个字与当前字。

从训练时间和小规模测试的结果考虑，本文所有实验字向量的维度均为 50。

5.1 实验设置

在有监督学习部分，本文使用的语料为 SIGHAN 2005 bakeoff 的分词语料。选取其中北京大学标注的数据用于训练、验证和测试。

原始语料只包含了训练集与测试集，在实验前，我们将原始语料的训练集前 90%当作我们自己的训练集，最后 10%当作开发集。测试集保持不变。最后训练集共有 1626187 个字，验证集包含了 160898 个字，测试集有 168973 字。

在非监督实验中，我们使用了两个语料，第一个语料（实验中称“小语料”）直接采用了北大标注的数据中的训练集，共 179 万字。第二个语料（实验中称“大语料”）在第一个语料的基础上，加入了搜狗新闻语料的精简版，其中涉及教育、文化、军事等一共 10 个类型的新闻语料。删除其中有乱码的句子后，最后得到的语料一共有 2723 万字。

非监督训练中，需要确定一个字典，字典选从大语料中出现的 1 万多个字中，选取出现

次数大于等于 5 次的所有字。剩下的字全都使用“unknown”特殊标记替代。一共 5449 字。

实验中，所有的最大熵模型均使用 liblinear 工具包计算。而神经网络实验由自己编写的代码完成，在训练集上训练，当开发集准确率达到最大值时，停止训练，取该模型用于测试。

5.2 基准实验

基准实验使用分词中较为常用的最大熵模型，特征选用一元及二元特征。

对于字 c_k，其特征向量具体包括：

- 一元特征 c_i，其中 i 为{k-2, k-1, k, k+1, k+2}，如果 c_i 超出了句子的边界，则使用一个特殊的符号“padding”来代替。
- 二元特征 $c_i c_{i+1}$，其中 i 为{k-2, k-1, k, k+1}，如果 c_i 或 c_{i+1} 超出了句子的边界，则忽略这个特征。

以上所有特征的权重均为 1。

基准实验一共有两个，第一个实验只使用了上述的一元特征，在后文中称作“最大熵一元特征”；第二个实验同时使用了一元特征和二元特征，在后文中称作“最大熵二元特征”。

以上两个基准实验均使用最大熵算法进行训练和测试。对字标签进行预测后，使用 viterbi 算法搜索最优路径。

为了展示神经网络模型以及字表示对于实验的影响，本文设计了多组对比实验：

1. 监督网络：使用第 3 章中所述的网络结构进行监督分词，其初始值选用均匀分布的随机数。
2. 监督网络+小语料字向量：在上一个实验的基础上，使用第 4 章描述的方法在 179 万字的小语料上训练得到的字向量作为初始值，训练分词网络。
3. 监督网络+大语料字向量：字向量使用 2723 万字的大语料训练得到，其余同上一个实验。
4. 大语料字向量最大熵：使用大语料训练生成的字向量作为特征，使用最大熵算法训练字标注器。在实验中，本文设定窗口大小为 5，字向量的维度为 50，因此每个字均有 250 个特征，各特征的权重对应窗口中每个字字向量的各维分量。
5. 随机字向量最大熵：将每个字的字向量替换成 50 个随机数，重复上一个实验。
6. 最大熵二元特征+字向量：使用大语料训练生成的字向量作为额外特征加入到“最大熵二元特征”实验中。即每个字的特征为一元特征、二元特征以及 250 个字向量特征。

5.3 实验结果及分析

实验结果如下：

编号	实验方案	准确率	召回率	F 值	未登录词召回率
	Sighan2005 最佳成绩（封闭）	0.946	0.953	0.950	0.636
	Sighan2005 最佳成绩（开放）	0.968	0.969	0.969	0.838
#1	最大熵一元特征	0.868	0.866	0.867	0.762
#2	最大熵二元特征	0.953	0.945	0.949	0.825
#3	监督网络	0.933	0.928	0.931	0.773
#4	监督网络+小语料字向量	0.942	0.934	0.938	0.792
#5	监督网络+大语料字向量	0.942	0.938	0.940	0.785
#6	大语料字向量最大熵	0.772	0.785	0.778	0.607
#7	随机字向量最大熵	0.559	0.590	0.574	0.498
#8	最大熵二元特征+字向量	0.953	0.946	0.949	0.815

表 1 实验结果

表 1 中列举了本文所做的一共 8 组实验。其中 #1 和 #2 为上一节中描述的两个基准实验。#3 到 #8 依次为上一节中描述的各个实验。。

#1 和 #2 为传统的最大熵分词方法得到的结果，与前人论文得到的结果相同，使用最大熵模型配合二元特征可以取得非常好的效果，该方法在 Sighan2005 的评测中，可以排到第三名。

#3 和 #6 相比，有巨大的优势，这里主要有两点原因：一、当特征数较少时，非线性的神经网络相比线性的最大熵模型有优势（与之相对的，如果特征数很多时，如 #2 中使用的二元特征，非线性模型无论是训练时间还是测试时间都会非常长）；二、神经网络模型在反向传播时，可以修改词向量，这相比直接把词向量作为输入特征的最大熵模型更为灵活。

#3、#4、#5 的比较中可以看出，无监督训练得到的字向量在作为有监督训练初始值时，可以显著地提升有监督学习的效果。其中 #4 虽然采用了无监督的数据进行训练，但实际上训练数据来自北大标注语料，因此可以看作是封闭训练的结果。#5 只能看作开放训练的结果。

#2 和 #8 中可以看出，将字向量作为附加特征辅助最大熵模型，效果几乎没有提升（只在小数点后第四位略有提升）。

#6 和 #7 的对比实验。值得注意的是，即使使用随机数来描述一个字，也可以取得超过纯猜测的效果（不到 0.25 的准确率）

字向量除了通过在有监督学习中看出其效果之外，可以直接通过字之间的相似度，看出其效果。表 2 展示了有监督学习得到的字向量，以及不同大小语料无监督学习得到的字向量的比较。这里选取了“一”、“李”、“江”、“急”这 4 个字。从对比中，可以很明显的看出，无监督方式学习得到的相似字，与原字在深层语义上更为相关。而且语料越大，这个效果越明显。

一			李			江			急		
监督	小	大	监督	小	大	监督	小	大	监督	小	大
八	四	三	沱	刘	杨	东	海	河	湃	僧	遥
三	两	两	尸	王	刘	哄	乡	山	瘫	尾	婉
六	各	六	玄	徐	赵	鲁	河	龙	泓	刨	擒
铵	三	二	孙	龚	郑	嘀	村	云	濮	除	互
二	五	四	郭	赵	郭	锐	县	海	潼	或	迅
七	那	五	券	袁	吴	*	山	城	悼	谁	凶
跌	这	num	赵	郭	邱	哨	湖	东	嘤	未	迫
四	几	几	袁	胡	冯	袅	岩	南	奕	微	损
吼	八	每	苔	潘	彭	辐	水	湘	慑	腕	后

表 2 各字向量得到的“一”、“李”、“江”、“急”的最相似的字

事实上，神经网络在训练时对初始值及各个参数是十分敏感的，包括随机梯度下降法中使用的学习速率，都会对结果造成影响。本实验中借鉴了[20]的方法，使用固定的学习速率，各层的学习速率与该层输入节点数的平方根成反比。可能换用其它的参数，可以获得更好的训练效果，甚至超过最大熵模型的效果。但是由于时间有限，本文并不能尝试各种不同的优化方案。同样的，对于无监督阶段，如果采用更大的语料，更充分的训练，也应当能取得更显著的效果。

本文得出如下结论：字向量的表示是一种较好的特征，使用字向量配合神经网络实现的

分词，相比一元特征有较大的优势。但是这种方法目前还不能取代人工设计特征，即使是简单的二元特征。随着数据量的增大，无监督学习得到的字向量也会越来越实用，相信使用更丰富的无监督训练语料，可以得到更有用的字向量。

6 总结与展望

本文探索了一种基于表示学习的中文分词方法。我们首先在大规模中文语料中学习字的语义向量表示，然后将学得的语义向量应用于有监督的中文分词。实验表明，表示学习是一种有效的中文分词方法，并在该领域展现出一定的潜力。然而我们发现，它尚不能取代传统基于人工设计特征的有监督机器学习方法。对表示学习方法的改进包括修改神经网络结构、修改网络的目标函数和使用更好的优化算法等。相信通过不断的改进，表示学习算法可以成为一种较实用的中文分词方法。

7 致谢

本文受国家自然科学基金项目（61070106，61272332，61202329），国家高技术研究发展计划（863 计划）（2012AA011102），国家重点基础研究发展计划（973 计划）（2012CB316300）以及网络文化与数字传播北京市重点实验室开放课题（ICDD201201）的资助。

参考文献

[1]《汉语信息处理词汇 01 部分: 基本术语（GB12200.1-90）6》，中国标准出版社, 1991。

[2] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Nature, 313(5786):504–507, 2006.

[3] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. J. Mach. Learn. Res., 999888:2493–2537, Nov. 2011.

[5] N. Xue. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 8(1):29–48, 2003.

[6] 刘群 基于层叠隐马模型的汉语词法分析

[7] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the 20th international conference on Computational Linguistics, COLING ’04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[8] B. Tang, X. Wang, and X. Wang. Chinese word segmentation based on large margin methods. International Journal of Asian Language Processing, 19(1):55–68, 2009.

[9] H. Zhao, C.-N. Huang, M. Li, and B.-L. Lu. Effective tag set selection in chinese word segmentation via conditional random field modeling. In Proceedings of PACLIC-2006, pages 87–94, 2006.

[10] K. Wang, C. Zong, and K.-Y. Su. A character-based joint model for chinese word segmentation. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING ’10, pages 1173–1181, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] Zhao H. and C. Kit. 2011. Integrating unsupervised and supervised word segmentation: The role of goodness measures. Information Sciences, 181(1):163 – 183, 2011.
[12] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 26th International Conference on Machine Learning, ICML, 2011.
[13] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. NIPS, 2011.
[14] B. Antoine, G. Xavier, W. Jason, and B. Yoshua. Joint learning of words and meaning representations for open-text semantic parsing. In Proc. of the 15th Intern. Conf. on Artif. Intel. and Stat., 2012.
[15] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 151–161, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
[16] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Soulie and J. H ´ erault, editors, Neurocomputing: Algorithms, Architectures and Applications, pages 227–236. NATO ASI Series, 1990.
[17] Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. NIPS (pp. 1081–1088).
[18] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 384–394, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
[19] 黄昌宁，赵海， 中文分词十年回顾
[20] D. C. Plaut and G. E. Hinton. Learning sets of filters using back-propagation. Computer Speech and Language, 2:35–61, 1987.