基于 DNN 的汉语框架识别研究*

赵红燕[1,2], 李茹[1,3], 张晟[1], 张力文[1]

(1. 山西大学 计算机与信息技术学院，山西 太原 030006;

2. 太原科技大学 计算机科学与技术学院，山西 太原 030024;

3. 山西大学计算智能与中文信息处理教育部重点实验室，山西 太原 030006)

摘要：框架识别是语义角色标注的基本任务,它是根据目标词激起的语义场景,为其分配一个合适的语义框架。目前框架识别的研究主要是基于统计机器学习方法，把它看作多分类问题，框架识别的性能主要依赖于人工选择的特征。然而，人工选择特征的有效性和完备性无法保证。深度神经网络自动学习特征的能力，为我们提供了新思路。本文探索了利用深度神经网络自动学习目标词上下文特征，建立了一种新的通用的框架识别模型，在汉语框架网和《人民日报》2003 年 3 月新闻语料上分别取得了 79.64%和 78.58%的准确率，实验证明该模型具有较好的泛化能力。

关键词：汉语框架；框架识别；深度神经网络；分布式表征

中图分类号：TP391　　**文献标识码：**A

Chinese Frame Identification with Deep Neural Network

ZHAO Hongyan[1,2], LI Ru[1,3],ZHANG Sheng[1],ZHANG Liwen[1]

(1.School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006,China;

2.School of Computer Science & technology, Taiyuan University of Science and Technology,

Taiyuan, Shanxi 030024 , China;

3.Key Laboratory of Ministry of Education for Computation Intelligence & Chinese Information Processing,

Taiyuan, Shanxi 030006,China)

Abstract: Frame identification is the basic task of semantic role labeling, which assigns a correct frame to the labeled target word based on the semantic scene. At present, the state-of-the-art methods used for frame identification are primarily based on statistical machine learning, and their performance of the identification strongly depends on the quality of the extracted features. However, the validity and completeness of the artificial selection feature can not be guaranteed. The ability of automatic learning characteristics of deep neural network provides us a new way of thinking. This paper explored the automatic learning characteristics of the target word context by deep neural network , established a new general frame identifying model, achieves a 79.64% and 78.58% accuracy on the Chinese FrameNet and the "people's Daily" news test corpus in March 2003. Experiment results show that the model has good generalization ability.

key words: Chinese FramNet; frame identification; deep neural network; distributed representation

1. 引言

语义角色标注（Semantic role labeling,简称 SRL）是浅层语义分析的一种有效方式。自 2004 年以来一直受到国内外自然语言处理学者的关注。汉语框架语义角色标注是基于汉语框架网（Chinese FrameNet，简称 CFN）语料资源的论元角色标注，旨在研究目标词激起的特定语义场景下的角色标注问题。语义角色标注技术在大规模语义知识库的构建、机器翻译、信息提取、自动文摘、智能问答、信息检索等应用领域都有着广泛的应用，其深入的研究对自然语言处理技术的整体发展有着重要意义[1]。

*收稿日期：　　定稿日期：

基金项目： 国家自然科学基金项目 (No.61373082)；国家 863 计划项目（No.2015AA015407）；山西省科技基础条件平台建设项目(No.2014091004-0103)；山西省回国留学人员科研资助项目(No.2013-015)；中国民航大学信息安全测评中心开放课题基金项目（No.CAAC-ISECCA-201402）

作为汉语框架语义角色标注的任务之一的框架识别包括未登录词元框架识别和歧义词元框架识别。其中，未登录词元框架识别旨在研究如何为能够激起 CFN 中的语义场景，但没有被收录到相应框架下的词元分配正确的语义框架。然而，歧义词元框架识别旨在研究如何为 CFN 中能够激起多个框架的词元分配一个正确的框架。对于未登录词元的框架识别，目前研究主要借助 WordNet,Wikipedia 和 VerbNet 等语义资源，通过相似度计算或者提取特征，利用统计机器学习方法建立分类器实现未登录词元的框架识别。针对歧义词元框架识别研究，采用的方法大多是借鉴“词义消歧”思想，利用已有句法分析等工具，人工建立特征，利用条件随机场(Conditional Random Field，简称 CRF)、支持向量机(Support vector machine，简称 SVM)、最大熵（maximum entropy,简称 ME）等分类器建立模型，把框架识别看作多分类问题[2,3,4,5]。

以上研究在框架识别任务上已经取得了一定的成效,但框架识别的性能主要依赖于人工选择的特征和现有的自然语言处理系统。一方面手工选择特征，费时费力，无法保证所选特征的有效性和完备性；另一方面现有自然语言处理工具中的误差传播也会影响框架识别的性能。并且现有框架识别研究大都是针对以上两个任务中的一个进行研究，不能实现对任意给定的目标词分配框架。

深度神经网络具有自动学习特征的能力，只需要给它提供一个底层的初始向量表征，通过网络自动学习学出更高级别的特征，给我们进行框架识别提供了新的思路。在此基础上，Hermann[6]提出基于分布式表征的框架识别方法，在 FrameNet 语料上取得较好的结果，在中文上还没有相关研究。

本文结合 Hermann 提出目标词上下文的分布式表征和深度神经网络方法建立一个通用的汉语框架识别模型，克服了传统利用统计学习方法选择特征和利用已有自然语言处理工具误差传播的弊端,实现了为任何给定的目标词分配合适的语义框架。

2.相关工作

2.1 汉语框架网

汉语框架网[7](Chinese FrameNet,CFN)是在刘开瑛教授的指导下，由山西大学从 2004 年开始建立。CFN 是一个以 Fillmore[8]的框架语义学为理论指导、 以伯克利的 FrameNet 为参照、 以汉语语料为依据的汉语词汇语义知识库。汉语框架网由词元库、框架库和例句库三部分组成。词元，是能够激起语义场景的词，也叫目标词。汉语框架是存储在人类经验中的图式化情境，既可以是一个实体，也可以是一种行为模式，甚至是一些社会习俗制度等。框架元素是语义场景中的各种参与者，包括核心框架元素、非核心框架元素和通用非核心框架元素三类。核心框架元素是框架语义场景中的必有成分。非核心框架元素表示目的、原因、时间等外围语义成分。核心和非核心框架元素因框架不同而不同。通用非核心框架元素作为框架库的补充，各个框架都适用。目前 CFN 已入库框架 361 个、词元 4547 个、标注例句 40000+条。据统计，CFN 中能够激起多个框架的目标词达到 1245 个，占总词元的 27.5%。因此框架识别是框架语义角色标注任务最基本但又重要的一步，它对框架语义角色标注任务有着直接的影响。

2.2 框架识别

框架识别作为 2007 年 SemEval 中框架语义分析的一个子任务被提出，包括未登录词元框架识别和歧义词元框架识别。

未登录词元框架识别主要借助 wordnet、verbnet 和 wikipedia 等语义知识库实现此任务。Aljoscha Burchardt 等[9]于 2005 年提出一种基于规则的未登录词元框架识别系统，

利用 wordnet 语义知识库为框架库中的词元选择一个 wordnet 词义，计算未登录词元和候选框架中词元的相似度，把未登录词元分配给相似度最大词元所在的框架，获得 39%的框架识别准确率。2007 年 LTH 研究小组[10]提出基于机器学习的未登录词元框架识别方法，选取 wordnet 的上下位关系作为特征，利用 SVM 构建分类器，取得 75.8%的框架识别准确率。MPennacchiotti 等[11]提出结合分布式模型与 wordnet 知识库的未登录词元框架识别模型，使框架识别准确率和召回率得到权衡。DipanjanDas 等[12]未借助任何语义资源，采用基于图的半监督学习方法，获得未登录词元 62.35%的准确率。陈雪丽等[13]2010 年利用哈工大同义词林，提出基于平均语义相似度计算及最大熵模型两种方法，采用静态特征和动态特征相结合的特征选择方法在 CFN 语料上和真实新闻语料上都取得了较好的效果。

歧义词元框架识别主要借助“词义消歧”思想，人工选择特征，采用 CRF、ME、SVM 等建立分类器进行实现。Cosmin Adrian Bejan 等[14]选择了 FrameNet 中 556 个歧义词元，每个词元至少包括 5 条例句，使用了 SVM 和 Maximum Entropy 为每个有歧义的目标词构造了一个多分类器进行框架排歧， 在 SVM 分类器上取得了 76. 71%的准确率。Richard Johansson 和 Pierre Nugues[15]针对歧义词元采用词形、目标词的词根、目标词依存关系集合和父节点、子节点等特征，利用 SVM 对每个歧义词元分别训练了一个分类器，针对 FrameNet 语料库中所有存在歧义的词元，取得了 84%的准确率。李茹[3]等提出基于依存分析的条件随机场模型进行汉语框架识别；李国臣[16]等研究了基于词元语义特征的汉语框架语义排歧方法，提出采用自动特征选择方法进行框架排歧。

3. 汉语框架识别模型

汉语框架识别是针对一个给定目标词句子，计算机能够根据目标词的上下文语境,在汉语框架库中自动给它选择一个合适的框架。其形式化描述如下：

$$f = \arg\max_{f_i \in F} p(f_i \mid w_t, C) \qquad (1)$$

其中 w_t 是目标词，f_i 是框架库中的第 i 个框架，C 是目标词的上下文集合，F 是框架集合。

3.1 汉语框架识别 DNN 架构

图 1 是我们进行汉语框架识别的 DNN 架构图，这个网络针对一个给定目标词的句子，通过 DNN 学习更抽象的目标词上下文特征,来实现框架识别。该网络主要包括上下文分布式表征层（输入层）、两个更高级别的特征学习层（隐层）及输出层。网络的输入层是基于依存关系抽取的上下文分布式表征，特征抽取过程在 3.2 节介绍。通过两个隐层学习目标词上下文的更好的表征，最后把学好的表征输入到一个 $soft\max$ 分类器。该分类器的输出是一个向量，向量的每个维度上的值表示当前目标词属于相应框架的概率，最后把概率最大的框架作为预测框架。

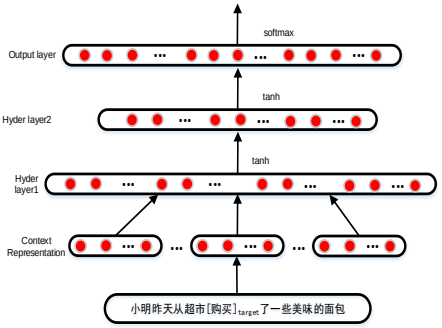


图 1 汉语框架识别深度神经网络架构

3.2 上下文特征抽取

要用深度神经网络来进行上下文特征学习，首先要对上下文进行分布式表示，把输入的

句子变成可以计算的实值，也就是寻找一个上下文表征函数 $g(x)$ 来表示 C 。原则上 $g(x)$ 可以是任何特征函数，但在这里我们考虑两种因素。一个是目标词直接或间接支配的依存关系作为插槽，形成特征模板，另一个是以依存关系对应的词向量去填充插槽，得到目标词的上下文初始表征，如果输入句子中不存在某种关系或该关系对应的词向量在 word embedding 库中不存在，则用 0 向量表示。

形式化描述，假设 x 是一个被标记了目标词 w_t 的句子， $g(x)$ 是 w_t 的上下文映射函数。如果词向量是 n 维，则 $g(x)$ 是句子 x 到 R^{nk} 的一个映射， k 是目标词支配的上下文依存关系类型数。例如，“他买了一本书”。如果 g 只考虑主胃关系（SBV）和动补关系（CMP），那么 $g:x\rightarrow R^{2n}$，前 n 维是主语“他”对应的词向量，由于本句中没有 CMP 关系，所以 $n\sim 2n$ 维都是 0，可表示为：

g(x)=[前 n 维是“他”对应的词向量，0，0，0...,0]

图 2 给出了例 1 的依存分析图①和上下文特征抽取过程②③④，带阴影的小圆圈表示对应的词向量。图 3 给出了例 1 的汉语框架语义角色标注结果。

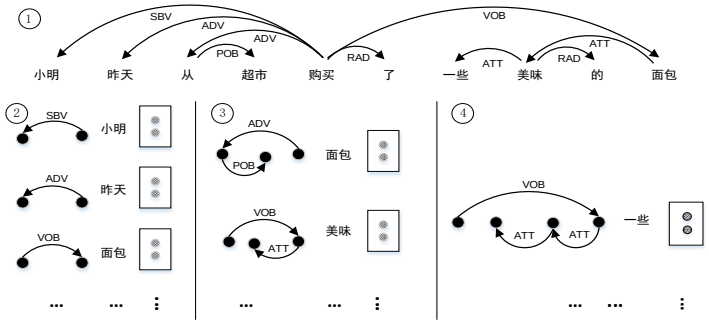


图 2 例 1 依存分析树与分布式表示特征抽取过程

例 1： 小明昨天从超市购买了一些美味的面包。

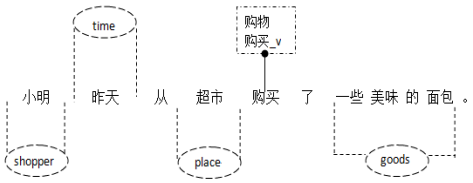


图 3 例 1 汉语框架语义角色标注结果

3.2.1 直接依存特征

从图 2 和图 3 可以看出和目标词有直接依存关系的块往往是目标词的核心框架元素或非核心框架元素，对目标词所属框架的判断有着直接的关系。我们首先考虑和目标词有直接依存关系的位置，如图 2②所示，和目标词“购买”有直接依存路径的有 SBV,ADV,VOB 等。如果只考虑直接依存关系的话,句子通过 $g(x)$ 到 R^{nk} 的映射，这里 k 就是直接依存关系的类型数，也是上下文模板的插槽数，然后根据依存关系找到对应的词元，用词元在 word embedding 中词向量来填充插槽，作为目标词上下文的表征，记为 T1。在这我们采用哈工大 Ltp 平台[16]进行依存分析，考虑了 14 种直接依存关系，如图 4 所示。

直接依存关系 T1(14 种)	ADV WP VOB SBV COO RAD CMP HED ATT FOB DBL LAD POB IOB

图 4 直接依存关系

3.2.2 间接依存特征

从图 2①和图 3 上看，除了和目标词有直接依存关系的句法成分外，有着间接关系的“超市”等也是“购买”框架元素。如果只考虑目标词的直接依存成分就会丢失很多有用的信息，为了获取更多的上下文信息，我们把和目标词有二级三级路径的句法成分也当作目标词的上下文特征，二级三级特征抽取过程如图 2③④所示。在 *CFN* 所有语料库中统计，和目标词有二级关系的有 110 种，有三级关系的有 497 种。由于二级三级关系太多，为了避免数据稀疏，论文中均选择二级三级关系的 *top30*。如图 5 所示。

二级依存关系 T2（30种）	VOB-ATT SBV-ATT COO-ADV ADV-POB COO-VOB HED-WP ADV-WP VOB-VOB VOB-ADV HED-ADV COO-SBV VOB-COO COO-COO COO-WP VOB-SBV HED-SBV ADV-ATT VOB-WP HED-VOB HED-COO COO-RAD CMP-POB SBV-WP COO-CMP ADV-ADV ATT-RAD ADV-RAD SBV-COO VOB-RAD ATT-ATT
三级依存关系 T3（30种）	VOB-ATT-RAD ADV-POB-ATT VOB-ATT-ATT COO-VOB-ATT SBV-ATT-RAD SBV-ATT-ATT VOB-VOB-ATT COO-COO-ADV COO-ADV-POB COO-COO-VOB HED-COO-ADV COO-SBV-ATT VOB-SBV-ATT HED-COO-VOB HED-SBV-ATT COO-VOB-ADV COO-VOB-VOB CMP-POB-ATT VOB-COO-ADV HED-VOB-ATT VOB-COO-VOB VOB-ATT-COO VOB-COO-WP COO-VOB-COO HED-ADV-POB VOB-ATT-ADV VOB-COO-LAD VOB-ADV-POB COO-COO-COO HED-COO-COO

图 5 二级和三级依存路径

把以上两种特征对应的词向量（图 2 带阴影的小圆圈）连接起来生成一个向量，来表示目标词的上下文，即 $g(x)$，作为框架识别神经网络的输入。

3.3 汉语框架识别网络学习

为了自动学习更好的上下文特征，我们设计了一个包含两个隐层的神经网络模型，如图 1 所示，学习过程包括前馈计算和反向传播两个阶段。两个隐层的激活函数都采用 tanh 函数。因为 tanh 导数具有以下特性:

$$\frac{d}{d_x}\tanh = 1 - \tanh^2 \quad (2)$$

该特性使得它在进行反向传播时计算梯度更容易。

3.3.1 DNN 前馈计算

网络的输入层为3.2节中抽取的上下文表征 $g(x)$，$g(x) \in R^{nk\times 1}$，n 是词向量的维度，k 是考虑的直接依存及二级、三级依存路径的关系的种类。把各种关系对应的词向量连接起来作为 DNN 网络的输入。

网络的第一个隐层（Hider layer1），有 n_1 个神经元，该层的输入为：

$$z_i^{(1)} = H_1 g(x) + b_1 \quad (3)$$

Hider layer1 输出为：

$$z_o^{(1)} = \tanh\left(z_i^{(1)}\right) \quad (4)$$

网络的第二个隐层(Hyder layer2)，有 n_2 个神经元，该层输入为：

$$z_i^{(2)} = H_2 z_o^{(1)} + b_2 \quad (5)$$

Hyder layer2 输出为：

$$z_o^{(2)} = tanh\left(z_i^{(2)}\right) \quad (6)$$

输出层(output layer)表示为：$y = U\tanh(z_o^2) + b_3$ (7)

其中 H_1, H_2, U 分别是第一个隐层、第二个隐层、输出层的权值矩阵，b_1, b_2, b_3 分别是第一个隐层、第二个隐层、输出层的阈值矩阵，初始的 $H_1, H_2, U, b_1, b_2, b_3$ 随机产生。输出层的神经元个数为 n_3 个，等于框架识别系统中框架的数量。用 $\theta = (H_1, H_2, U, b_1, b_2, b_3)$ 表示深度神经网络中的所有参数，则 y 是 θ 的函数，$y \in R^{n_3\times 1}$，输出层的每个节点 y_i 表示目标词在它的上下文中属于第 i 个框架的未归一化 log 概率。最后使用 $soft\max$ 激活函数将输出值 y 归一化成概率：

$$p(f_i \mid x,\theta) = \frac{e^{y_i}}{\sum_{j=1}^{n_3} e^{y_j}} \qquad (8)$$

3.3.2 DNN 反向传播训练

模型训练的过程就是要通过已经标注的训练样本$(x^{(i)}, f^{(i)}), i \in N$（N 为训练样本数，$x^{(i)}$是第$i$个训练样本，$f^{(i)}$是第$i$条句子标注目标词所属框架），寻找参数集合$\theta$使得带正则项的对数似然概率最大化，似然函数如下：

$$L(\theta) = \sum_{i=1}^{N} \log p(f^{(i)} \mid x^{(i)},\theta) + R(\theta) \quad (9)$$

$R(\theta)$是为了防止过拟合加的正则项。我们采用随机梯度上升方法学习似然函数中的参数θ。在 DNN 不同层之间采用反向传播算法，不断进行迭代，更新参数，直到达到预设精度或最大迭代次数，迭公式如下：

$$\theta \leftarrow \theta + \eta \frac{\partial \log p(f \mid x,\theta)}{\partial \theta} \qquad (10)$$

η是学习率。

4. 实验设置和结果分析

在这一部分我们给出实验所用的数据集、评价指标、实验的参数设置及实验所取得的结果，及和其它模型的比较情况。

4.1 数据集

本文的训练集(称为 train)选用 CFN 例句库中 25000 条句子，共涉及 1567 个词元，180 个框架。测试集分为三部分，测试集 1（称为 test1）选用 CFN 中未出现在训练集中的 5000 条句子；测试集 2（称为 test2）选用《人民日报》2003 年 3 月的 986 篇新闻，共 9573 条句子，去掉不能激起语义场景的句子后，选择了 10367 个目标词作为候选目标词；测试集 3（称为 test3）采用李等[3]2010 年 Coling 会议上所用数据集，该数据集包括“表示”，“想”，“叫”，“有”，“倒”，“下降”，“装载”七个歧义词元，128 条句子作为测试数据，每条句子人工标注目标词。

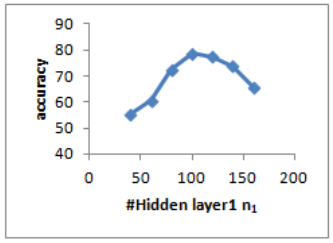
预处理：本实验中所有训练数据和测试数据均利用哈尔滨工业大学语言技术平台 LTP[17]进行依存分析。

4.2 评价指标

实验中评价指标采用:

$$accuracy = \frac{\sum_{i=1}^{T}\sum_{j=1}^{5} v_{ij}}{\sum_{i=1}^{T}\sum_{j=1}^{5} n_{ij}} \times 100\% \quad (11)$$

T是测试语料目标词的个数，v_{ij}是第 j 次交叉验证目标词t_i分类正确的句子数，n_{ij}是第 j 次交叉验证含有目标词t_i的测试样本数。



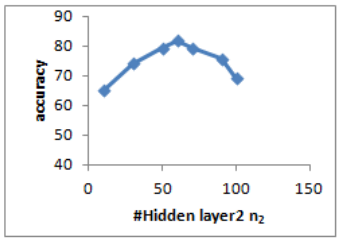


图 6 隐层参数的影响

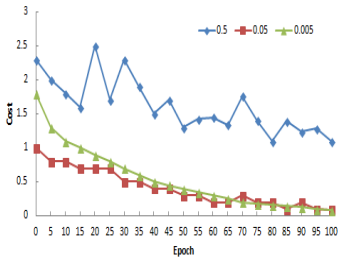


图 7 学习率影响

4.3 参数设置

在实验中我们利用五折交叉验证的方法，调整提出模型中的三个超参数，即 Hider layer1 的神经元个数n_1，Hider layer2 的神经元个数n_2，及学习率η。隐层神经元个数的调整方法是先根据经验分别给n_1、n_2赋一个初值，通过固定其中一个调整另一个，直到准确率不再提升为止。图 6 中给出了n_1,n_2和 accuracy 的变化关系，可以看到 Hidden layer1 神经元个数在 100 附近准确率不再增长，Hidden layer2 神经元个数在 60 左右准确率 accuracy 有所下降，并且从图 6 可以看到，在n_1=100，n_2=60，准确率也达到最大。学习率η，通过实验（0.005，0.05，0.5）三种取值，如图 7 所示。实验迭代了 100 次，当η=0.5 时，代价函数始终都震荡的很明显，这是由于我们使用了随机梯度进行迭代时，由于学习率太大，使算法在学习过程中越过了最小值。当η=0.05 时代价开始下降很快，但在大约 70 次迭代后有轻微震荡，此时容易跨过全局最小值达到局部最小值。当η=0.005 时，代价函数一直平滑下降到谷底，因此综合考虑迭代速度和代价函数后，本实验选择学习率为 0.005。表 1 给出了我们实验中所用的超参取值。

表 1 实验中的超参取值

Hyper parameter	Hidden layer1(n_1)	Hidden layer2(n_2)	Learning rate(η)
value	100	60	0.005

表 2 在 test1 上实验结果

feature	All(acc)	Ambiguous(acc)	Unseen(acc)
T1	67.21%	64.13%	53.72%
T1+T2	77.53%	72.4%	65.63%
T1+T2+T3	79.64%	74.37%	67.21%

4.4 实验结果及对比

本文的词向量采用北京理工大学的训练好的中文 word Embedding 库，规模约 30 万，每个词向量 100 维。首先对训练语料和测试语料利用哈工大的 LTP 平台进行依存分析，然后利用 3.2 节介绍的方法提取上下文特征。本文考虑 T1，T2，T3 三种特征，分别对所有目标词、Ambiguous(歧义)目标词、未登录(Unseen)目标词做了实验，结果如表 2 所示. feature 是采用的特征；All 是所选框架包含的所有目标词；Ambiguous 是框架中可以激起多个框架的歧义目标词；unseen 是框架库和标注语料中没有出现的目标词，即未登录目标词。

在我们的方法中，网络抽取了直接依存关系和二级三级路径的依存关系对应的词作为目标词的上下文，为了分析每种特征的有效性，我们采用了 ablation 实验，分别利用 T1、T1+T2、T1+T2+T3 组合在 test1 数据集上的实验结果如表 2 所示，结果表明和目标词有直接

依存关系的词对框架识别结果显著，在 test1 的所有测试数据上达到了 67.21%的准确率，在歧义目标词和未登录目标词的框架识别上也分别达到了 64.13%和 53.72%的准确率。分析其原因，发现直接依存关系往往是目标词的核心框架元素，而核心框架元素是一个框架在概念理解上的必有成分，在不同的框架中核心框架元素的类型和数量都不相同，核心框架元素显示出一个框架的个性，对框架的识别起着决定性的作用。在加入二级路径依存关系后，在所有测试的目标词中准确率提升了 10.32 个百分点,在歧义词元框架识别准确率提升了 8.27 个百分点，而在未登录目标词框架识别的准确率提升最多达到 11.91 个百分点，可见加入二级依存关系结果提升显著，尤其是在未登录词元的框架识别上。经分析发现二级依存关系大都是目标词的非核心框架元素，非核心框架元素表达目标词所激起语义场景的时间、空间、环境条件、原因、目的等外围语义成分，这些成分对于框架的识别有一定的促进作用。比如例 1 中的“超市”就是目标词“购买”的地点。三级依存关系加入后虽有提升，但不如二级显著，最高提升 2.11 个百分点，究其原因，三级依存关系要么是目标词的通用非核心框架元素，要么不是目标词的框架元素，而通用非核心框架元素在每个框架中承担的语义角色都一样，因此对目标词进行框架识别时区分度不大。

由于中文 CFN 起步较晚，没有公开的数据集，在汉语框架识别方面研究也不多。本文只能跟目前已有的研究做一个宏观的比较，本实验和已有汉语框架识别模型比较结果见表 3。表 3 中 Model 是框架识别所用的模型,其中 proposed 为本文提出的模型；target 是研究中选用的目标词数量。

表 3 本文提出模型与其他模型对比

Model	feature	target	All(acc)	Ambiguous(acc)	Unseen(acc)
SVM	Word+pos+命名体+父节点	23	-	77.62%	-
ME	词分布+位置	88	-	58.11%	-
T-CRF	Word+pos+tree(6 种树特征)	7	-	81.46%	-
Proposed	词分布	1567	79.64%	74.37%	67.21%

表 3 中给出了目前研究汉语框架排歧的一些模型与实验结果（文献 3，16，18），并与本文提出的方法做了一个比较（采用 T1+T2+T3 组合特征分布式表示，测试语料用 test1）。可以得出以下结论：

（1）使用传统的特征进行框架排歧时，特征越丰富，模型性能越好。但特征的选择依赖于人的经验和知识库，人是不可能选出最好的特征的。

（2）模型二可以看出利用词分布作为特征，通过最大熵模型并不能取得比传统方法好的结果。

（3）可以看出本文提出的以基于依存位置提取的上下文分布式表示，作为初始输入，通过 DNN 学习更好的特征表示对于框架识别是有效的。

（4）而且传统的模型都是选择极少数能激起多个框架的词元建立模型的，比如效果较好的模型三，仅选择汉语框架中的“表示、想、叫、有、倒、下降、装载”7 个词元做的实验，模型一中选择 23 个歧义词元，模型二中选择了 88 个有歧义的词元。这些模型不能实现对未登词元和所有目标词分配语义框架，不能直接应用到汉语框架语义角色标注任务中。

而本文提出的方法对目标词没有要求，可以实现对所有目标词进行框架识别任务。在本实验中涉及到目标词 1567 个，远远大于以上模型，在通用框架识别任务上达到了 79.64%的准确率，因此本文提出的模型更具有泛化能力。分析本模型对于有歧义的目标词识别略低的原因，发现框架库中有些目标词虽然能够激起多个框架，但在某些框架下并没有相应的标注例句，因此可以通过增加例句来提高有歧义目标词识别性能。未登录目标词的识别低的原因

除了语料库不足外，也存在有些词在词向量表中找不到的因素。因此，如果把 CFN 语料加入词向量训练，可能会提升框架识别的整体效果。

为了和 Li 等实验结果进行对比，本文也在 Li 等 2010 年 Coling 会议所用 test3 数据集上做了实验，本实验只是针对七个歧义目标词，实验过程和上述过程相同，实验结果如表 4 所示。本文提出方法比 Li 等所用方法在相同数据集上准确率提高了 4.23%，由此可见针对小规模语料，本文提出的基于目标词依存关系的上下文分布式表征的深度学习方法对目标词所属框架识别具有较好的效果。

表 4 在 test3 数据集实验结果与 Li 比较

model	准确率
Li	81.46%
Proposed	85.69%

另外，为了说明本模型的通用性，本文采用《人民日报》2003 年 3 月新闻语料作为测试集，对本文所提出模型进行测试。从表 3 可以得知总是 T1+T2+T3 取得最好结果，所以这里选用三个组合特征，在 test2 上实验结果如表 5 所示。

表 5 test2 实验结果

feture	All(acc)	Ambiguous(acc)	Unseen(acc)
T1+T2+T3	78.58	73.14	64.9

由表 5 可见，在开放数据集的所有数据上、歧义词元及未登录词元的框架识别准确率取得的实验结果均和 CFN 例句库中数据取得的结果相差不大，因此，本文提出的方法具有较好的通用性。

以上实验结果表明，本文提出的深度神经网络方法针对少量的歧义目标词进行框架识别结果优于传统的基于特征的统计模型，并且在开放语料上取得和 CFN 语料类似的准确率，说明通过深度神经网络可以学习到更好的特征，而有些特征人是无法捕捉到的。

5. 结论和展望

本文初步探索了 DNN 在 CFN 框架识别任务上的应用，实验表明本文提出的利用依存关系生成目标词的上下文分布式表征，通过 DNN 自动学习目标词上下文的更好的表征，有助于汉语框架的识别。本方法把传统的框架排歧、未登录目标词框架识别及框架识别任务统一在一个模型下，能够为汉语框架语义角色标注任务提供服务。为了评价本模型的性能，在《人民日报》新闻语料上进行了测试，取得了和 CFN 语料的结果相差不大。并且采用和 Li 等同样的数据集对七个歧义词元进行框架排歧，框架识别结果比 Li 等模型框架识别准确率提升了 4.23 个百分点。

关于下一步的工作，本文所提出方法，输入分布式表征维度较高，模型参数较多，学习过程计算量较大，下一步可以通过卷积神经网络和 Relu 激活函数来优化本模型；另一方面，把框架识别应用到语义角色标注任务中，实现汉语框架语义角色标注自动化。

参考文献：

[1]李济洪. 汉语框架语义角色的自动标注技术研究：[D]. 太原：山西大学博士学位论文，2010.

[2]Ken Litkowski. CLR: Integration of FrameNet in a Text Representation System[C] //Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, 2007:113-116.

[3]Ru Li, Haijing Liu, Shuanghong Li.Chinese Frame Identification using T-CRF Model[C] //Proceedings of International Conference on Computional Linguistics. Beijing, 2010: 674 -682.

[4]Cosmin Adrian Bejan, Hathaway Chris. UTD-SRL:A pipeline Architecture for Extracting Frame Semantic

Structures[C] //Proceedings of the 4th International Workshop on Semantic Evaluations.Prague, 2007:460-463.
[5] C.Baker, M.Ellsworth, and K.Erk.SemEval-2007 Task 19: Frame Semantic Structure Extraction[C]// Proceedings of the 4th International Workshop on Semantic Evaluations.Prague, 2007:99-104.
[6]Karl Moritz Hermann, Dipanjan Das, Jason Weston Kuzman Ganchev. Semantic Frame Identification with Distributed Word Representations[C]//Meeting of the Association for Computational Linguistics . Baltimore, USA. 2014:1448-1458.
[7]刘开瑛．汉语框架语义网（ＣＦＮ）构建现状[C]．//第四届 全国学生计算语言学研讨会会议论文集．2008:1-7.
[8]C. J. Fillmore. Frame Semantics[J]. In Linguistics in the Moring Calm, Hanshin Publishing Co.. Seoul, South Korea. 1982: 111-137.
[9] Burchardt A, Erk K, Frank A. A WordNet detour to FrameNet[C]//Proceedings of the GLDV 2005 Germa-Net II Workshop Bonn , Germany,2005.
[10]R Johansson, P Nugues.Using WordNet to extend FrameNet coverage[C]//Proceedings of the workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages.Tartu,2007.
[11]M Pennacchiotti, DDe Cao, R Basili, D Croce,et al.Automatic induction of FrameNet lexical units[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu,2008:457-465.
[12]Dipanjan Das, Noah A Smith. Semi-Supervised Frame-Semantic Parsing for Unknown Predicate[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, 2011:1435-1444.
[13]陈雪丽,李茹,王赛等.汉语框架网中未登录词元的框架选择[J].中文信息学报.2014,28(3):48-54,61.
[14]Cosmin Adrian Bejan, Hathaway Chris ．UTD-SRL：A Pipeline Architecture for Extracting Frame Semantic Structures[C]．//In 45th annual meeting of Association for Computational Linguistics, 2007:460-463.
[15]Richard Johansson, Nugues Pierre.LTH: Semantic Structure Extraction using Nonprojective Dependency Trees[C] //Proceedings of the 4th International Work on Semantic Evaluations. Prague, 2007:227-230.
[16] 李国臣,张立凡,李茹等.基于词元语义特征的汉语框架排歧研究[J].中文文信息学报.2013,27(4):44-51.
[17] 哈尔滨工业大学 LTP 平台：[CP],http://www.ltp-cloud.com/document/#api_rest_note
[18] 党帅兵,李国臣,王瑞波等． 基于词分布表征的汉语框架排歧研究[J].中北大学学报.2015,36(3):328-332,337.

赵红燕（1977-），女，博士生，主要研究领域为中文信息处理。
Email:lrxzhy@163.com;

李茹（1963-），女，博士，教授，主要研究领域为中文信息处理与数据库技术。Email:liru@sxu.edu.cn 通讯作者；

张晟（1991-），男，学士，主要研究领域为中文信息处理。Email:zhangsheng20xy@163.com.

张力文（1991-），男，学士，主要研究领域为中文信息处理。Email: 505646231@qq.com.

作者联系方式：

赵红燕，山西大学，邮编：030006, 电话：13994210015，Email:lrxzhy@163.com。