文章编号：1003-0077（2011）00-0000-00

基于 Dropout 正则化的汉语框架语义角色识别*

王瑞波[1,2]，李济洪[1]，李国臣[3]，杨耀文[4]

（1.山西大学 软件学院，山西省 太原市 030006；

2. 山西大学 计算机与信息技术学院，山西省 太原市 030006；

3. 太原工业学院，山西省 太原市 030008；

4. 山西大学 数学科学学院，山西省 太原市 030006）

摘要：汉语框架语义角色识别是汉语框架语义分析的重要任务之一。本文基于汉语词语、词性等特征的分布式表示，使用一种多特征融合的神经网络结构来构建汉语框架语义角色识别模型。鉴于可用的训练语料规模有限，本文采用了 Dropout 正则化技术来改进神经网络的训练过程。实验结果表明，Dropout 正则化的加入有效地缓解了模型的过拟合现象，使得模型的 F 值有了近 7%的提高。本文进一步优化了学习率以及分布式表示的初始值，最终的汉语框架语义角色识别的 F 值达到 70.54%，较原有的最优结果提升 2%左右。

关键词：汉语框架网络；语义角色识别；Dropout 正则化；

中图分类号：TP391　　**文献标识码：**A

Semantic Role Identification of Chinese FrameNet based on Dropout Regularization

Wang Ruibo[1,2], Li Jihong[1], Li Guochen[3], Yang Yaowen[4]

(1. School of Software Shanxi University, Taiyuan, Shanxi 030006, China;

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

3. Taiyuan Institute of Technology, Taiyuan, Shanxi 030008, China;

4. School of Mathematic Sciences, Taiyuan, Shanxi 030006, China)

Abstract: Semantic role identification is an important task for semantic parsing of Chinese FrameNet. Based on distributed representations of Chinese word, part-of-speech and other symbolic features, we built our semantic role identification model by employing a kind of multi-feature-integrated neural network architecture. Due to the size of training corpus is relatively small, we adopted dropout regularization to improve quality of the training process. Experimental results illustrate that, dropout regularization can effectively alleviate over-fitting of our model, and the F-measure increases about 7%. We further optimized learning rate and initial values of word embedding. The final F-measure of our semantic role identification model achieve 70.54%, which is higher about 2% than the state-of-the-art result.

Key words: Chinese FrameNet; Semantic Role Identification; Dropout Regularization;

1 引言

语义角色标注是自然语言处理领域的一个重要的任务。给定一个句子中的目标词，语义角色标注任务的目标是自动识别该目标词所支配的所有语义角色并标注角色的类型。因此，高精度的语义角色标注模型为后续的句义分析和篇章理解奠定了重要基础，也为机器翻译、信息检索、自动文摘等应用系统提供语义上的支持。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（NNSFC-61503228）；NSFC-广东联合基金（第二期）超级计算科学应用研究专项资助”和“国家超级计算广州中心支持”；

作者简介：王瑞波（1985—），男，博士研究生，讲师，主要研究方向为自然语言处理；李济洪（1964—），男，教授，主要研究方向为统计自然语言处理；李国臣（1963—），男，教授，主要研究方向为中文信息处理；杨耀文（1990—），男，硕士研究生，主要研究方向为中文信息处理。

汉语框架网络是一种重要的汉语词语语义分析和理解的资源。它以框架语义学为背景，为词语的框架义分析以及后续的篇章理解提供了重要的理论依据以及形式化的方法。特别是，汉语框架网络不仅使用框架来体现词语的意义，并通过框架之间的依存关系来刻画词语之间的语义关系，还引入了核心依存图来构架一条汉语句子的句义骨架。从篇章的角度来看，Fillmore 针对英语在文献[1]中给出了使用框架语义学进行文本理解的一些重要理论分析成果。这非常值得汉语框架语义分析借鉴。然而，不管是汉语句子的句义构架还是汉语文本的语义理解，所面临的一个重要前提是构建一个高精度的汉语框架语义角色自动标注模型。

目前，汉语框架语义角色自动标注模型的精度仍然较低。李济洪等[2]以及宋毅君[3]等使用条件随机场模型，将词层面的特征以及句法块层面的特征融合到汉语框架语义角色自动标注模型中。实验结果表明，目前的模型的自动标注精度不到 70%(F 值)。他们的深入分析表明，汉语框架语义角色自动标注模型任务的最大难点在于语义角色识别的精度较低。因此，在后续的研究中，他们将汉语框架自动标注任务分割成语句角色的识别和分类两个子任务，并着重对语义角色的识别进行研究。

传统的一些汉语框架语义角色识别模型，主要是基于条件随机场，最大熵模型以及支撑向量机等机器学习算法进行构建。这些算法直接将词、词性等符号信息作为特征进行使用。它们忽略了这些特征之间的语义相关性。另外，由于词特征个数较多，导致模型的特征矩阵维度很高，容易导致模型的过拟合。

自 2006 年深度学习开始兴起，很多研究者开始使用深度神经网络的技术来解决图像识别、自然语言理解等领域的任务。对于自然语言处理来说，词语的分布式表示技术受到很多研究者的青睐[6,11]。词语分布式表示的目的是将词语、词性等语言学上的符号通过一些低维的实数向量来进行表示；然后，通过实数向量的代数运算，计算出词语等语言学符号之间的语义关系。针对词语的分布式表示信息，大量的学习模型被开发出来，例如，C&W 模型[6]等。

词语分布式表示信息的引入，为语义角色识别模型的构建提供了一个新的模式。Collobert 等[6]给出了一种基于神经网络的语义角色识别模型。该模型使用分布式表示信息将离散的词、词性等特征映射成低维的实数向量，并将该向量作为神经网络的输入。该模型很大程度上解决了传统模型中特征矩阵维度较高的问题。另外，该模型也有效地利用了分布式表示信息所蕴含的词语语义间的相关性。不过，当可用的训练语料较少时，神经网络中很多待估参数无法得到有效的估计，从而使语义角色识别模型产生严重的过拟合问题。

考虑到汉语框架语义角色识别的语料规模较小，结合神经网络模型的上述特点，我们使用 Dropout 正则化来改进神经网络算法的训练过程，进而缓解汉语框架语义角色识别模型的过拟合现象。在第 5 节，我们给出了使用 Dropout 正则化技术的识别模型。实验结果表明，加入 Dropout 正则化后，模型的识别性能有着显著的改善，F 值提升了近 7%。

本文的主要工作在于：针对汉语框架语义角色识别任务，基于一种多特征融合的神经网络结构来构建识别模型；然后，从 Dropout 正则化的训练方法、初始 Embedding 的设置以及学习率的选择等方法来优化模型的训练过程。最终的汉语框架语义角色识别模型的 F 值达到 70.54%，比宋毅君等[3]给出的最优识别性能高出仅 2%。

本文的组织结构如下：第 2 部分给出了汉语框架语义识别任务的介绍，并给出了分布式表示信息对模型性能改善的积极影响；第 3 部分提出了本文所使用的神经网络结构，并给出了 Dropout 正则化的训练方法；第 4 部分叙述了实验语料以及一些实验设置；第 5 部分给出了实验结果及相应的分析；最后，对全文进行了总结，并给出了进一步的研究方向。

2 汉语框架语义角色识别任务

汉语框架语义角色识别任务是指：给定一条汉语句子及目标词，在目标词的框架已知的条件下，从句子中自动识别出目标词所搭配的语义角色的边界。例如，对于汉语句子“英方

面作为报复措施也宣布 4 名俄罗斯大使馆的外交官为不受欢迎的人。”，针对目标词“宣布”，人工给出的框架语义角色标注如下:

<spkr 英方面 > 作为报复措施也 <tgt [陈述] 宣布> <msg 4 名俄罗斯大使馆的外交官为不受欢迎的人> 。

上述标注中，目标词“宣布”激起了“陈述”框架；在“陈述”框架所表达的语义场景中，“英方面”是说话者(spkr)，而“4 名俄罗斯大使馆的外交官为不受欢迎的人”为说话者所要传达的信息(msg)。在语义角色识别任务中，我们仅需要确定出“宣布”所搭配的语义角色为“英方面”和“4 名俄罗斯大使馆的外交官为不受欢迎的人”，不需要确定这两个语义角色的类型。

一般来说，在进行框架语义角色识别之前，我们会先对句子进行分词。由于中文的句法分析技术尚不成熟，本文仅考虑在词层面进行汉语框架语义角色识别的研究。对于一条已经分好词的汉语句子 $S = w_1, w_2, ..., w_n$，我们可以使用标记集合{I,O,B,E,S}将其对应的语义角色边界形式化成一个标记序列 $T = t_1, t_2, ..., t_n$，其中，w_i 和 t_i 分别为句子 S 中的第 i 个词及其边界标记，$t_i \in \{I, O, B, E, S\}$。语义角色识别问题通常被转化成如下的优化问题:

$$T^* = \arg\max_T P(T = t_1, ..., t_n \mid S = w_1, ..., w_n)$$

针对该优化问题，研究者通常使用各种统计机器学习算法来求解上述的条件概率。常用的机器学习算法有：条件随机场模型[2]，最大熵模型[4]，支持向量机[5]等。2006 年以后，随着深度学习技术的成熟，越来越多的研究者开始尝试使用神经网络模型来构建语义角色识别模型[6,7,8]。

基于神经网络构建的自然语言处理模型，通常会使用词、词性等语言学特征的分布式表示来作为神经网络的输入。也就是说，词、词性等特征通常不会直接参与模型的运算，而是通过一个表示矩阵来映射成一个实数向量。然后，神经网络算法基于该实数向量作为输入来建立词性标注、短语识别、语义角色识别等自然语言处理模型。Collobert 等[6]开发的自然语言处理模型便借助了词语、词性等的分布式表示信息。他们的实验结果表明，融合了分布式表示的自然语言处理模型的性能与之前的最优模型的性能是可比的。但是，目前并未有结论证明，融合分布式表示信息的自然语言处理模型的性能会有显著提高。

对于汉语框架角色识别模型，根据模型是否使用了分布式表示信息，我们可以将模型分为两大类：未使用分布式表示的识别模型和基于分布式表示的识别模型。在下面的两个小节中，我们分别阐述两类模型的特点以及当前的模型性能。

2．1 未使用分布式表示的识别模型

传统的汉语框架语义识别模型通常采用条件随机场、最大熵模型、支撑向量机等机器学习算法[2,3,4,5]。这些未使用分布表示的识别模型通常直接使用词、词性等的离散语言符号特征。这些符号特征从可用的语料资源中直接被抽取出来，然后根据机器学习算法的特征函数来转化成一个高维的 0-1 特征矩阵。

这种类型的识别模型往往具有两个主要缺点：第一、所形成的特征矩阵的维度较高；第二、特征之间的语义相关性无法得到体现。其中，第一个缺点是显然的。例如，在英文中，研究者通常抽取[-2,2]窗口内的词语作为特征。假设，英文中的词语为 30000 个，则通过这种方法抽取出的词语特征会达到 150000 个左右；如果研究者在加入一些词语之间的组合特征等，特征矩阵的维度将会指数级的增加。李济洪等基于条件随机场开发的汉语框架语义标注模型[2]中，训练语料仅有 6000 多句，但是模型中所抽取的特征函数却达到了近千万。这非常容易导致标注模型的过拟合，因此，李济洪等开发了一套正交表特征选择方法来调节所选特征的数量，并且使用 L2 正则化来缓解标注模型的过拟合现象。

对于第二个缺点，很多研究者试图将各种语义资源加入到标注模型中来刻画词、词性等之间的语义上的相关性。例如，Li 等试图将 FrameNet、WordNet 等语义资源集成起来改善

语义分析模型的性能[9]。邵艳秋等从中文概念词典中抽取出语义特征来改进语义角色标注模型[10]。尽管这些工作表明了语义资源对于语义角色标注模型的积极作用，但效果并不显著。

另外，对于语义角色识别任务，尽管已经有了 PropBank、CTB、FrameNet 等可用的语料库，但这些语料库的规模相对较小。如何使语义角色识别模型突破语料资源的限制，融合大量生语料资源，进而达到更高的模型性能，也是传统的语义识别模型所面临的重要问题。

在未使用分布式表示的汉语框架语义角色识别模型中，目前最好的结果由宋毅君等的工作给出[3]。他们使用条件随机场算法来构建汉语框架语义识别模型，该模型分别融合了 12 个词层面特征以及 15 个句法块层面特征。基于词层面的特征，语义角色识别模型的 F 值达到 68.51%；加入块层面特征后，模型的 F 值仅有 0.01 的提高，达到了 68.52%的 F 值。

2．2 基于分布式表示的识别模型

对于上一节中所给出的问题，一个有效的解决方案是：将词、词性等特征的分布式表示信息引入到汉语框架语义识别模型中。词、词性的分布式表示是将词、词性等符号特征映射成一个低维的实数向量。词、词性等特征间的语义关系可以通过它们对应的实数向量之间的代数运算计算出来。

基于词、词性的分布式表示信息，模型的特征矩阵的维数可以得到大幅度的压缩。例如，在上一节所给的例子，假设每个词语可以由 300 维的实数向量表示，[-2,2]窗口内的词语特征可以从原来的 150000 维压缩到 1500 维。特征矩阵得到了近 100 倍的压缩。另外，特征之间的语义相关性可以由实数向量之间的运算反映出来。

针对大量生语料的利用问题，Turian 等基于分布式表示信息给出了一般的半监督学习框架[11]。该框架认为，使用大量的未标注语料来训练出词、词性的分布式表示信息。这些分布式表示信息中凝聚了这些未标注语料中的语义信息。然后，神经网络算法直接利用该分布式表示信息作为输入来构建自然语言处理模型。这样，大量未标注语料的信息便可被有效利用。目前，多种词语的分布式表示学习算法被开发出来，例如，C&W 算法[6],Word2Vector 算法[12]，Glove 算法[13]。

目前，以后很多自然语言处理模型采用了词、词性等特征的分布式表示信息。例如，Collobert 等将维基百科和路透社的语料训练所得词语分布式表示信息，加入到神经网络模型中来构建词性标注、命名实体识别以及语义角色标注模型中[6]。李国臣等将字的分布式表示信息加入到神经网络中来自动识别汉语基本块[14]。

针对汉语框架语义角色识别模型，我们在第 3 节中给出了一种融合多种特征的神经网络结构。然后给出了基于 Dropout 正则化来训练该神经网络结构。

3 基于 Dropout 正则化的神经网络算法

本节中使用神经网络构建汉语框架语义角色识别模型。本文所构建的模型对不同词语对应的标记之间做出了独立性假设，即：

$$P(T=t_1,...,t_n \mid S=w_1,...,w_n)=\prod_{j=1}^{n} P(t_j \mid S=w_1,...,w_n)$$

进而，本文使用神经网络来预测每一个词的标记概率 $P(t_j \mid S=w_1,...,w_n)$，并选择概率最大的标记作为输出。在估计该标记概率时，我们从句子 S 中抽取出词、词性、位置和目标词四种特征，并使用这四种特征作为神经网络结构的输入。为了避免神经网络模型产生过拟合现象，本节将 Dropout 正则化技术引入到神经网络的训练方法中。

3．1 汉语框架语义识别的神经网络结构

本文使用图 1 所给出的神经网络结构来构建汉语框架语义角色识别模型。该神经网络在典型的三层神经网络结构上，将词、词性、位置和目标词四种特征进行融合。在该神经网络

结构中，四个 Embedding 矩阵存储了四种特征的分布式表示；词、词性、位置和目标词等离散特征通过 Embedding 矩阵来映射成相应的实数向量；然后，四种特征的实数向量被拉直拼接后作为输入；接着，Sigmoid 函数被用作对输入层进行非线性变换；最后，使用 softmax 函数来得到待标词语的标记概率。

图 1 中的神经网络结构与 Collobert 等所给的模型结构[6]的主要区别在于：本文的神经网络结构可以允许不同类型特征灵活选择特征窗口。但是，Collobert 给出的模型结构要求所有特征的窗口都相同。另外，本文使用的非线性函数为 Sigmoid，而不是 hardtan。

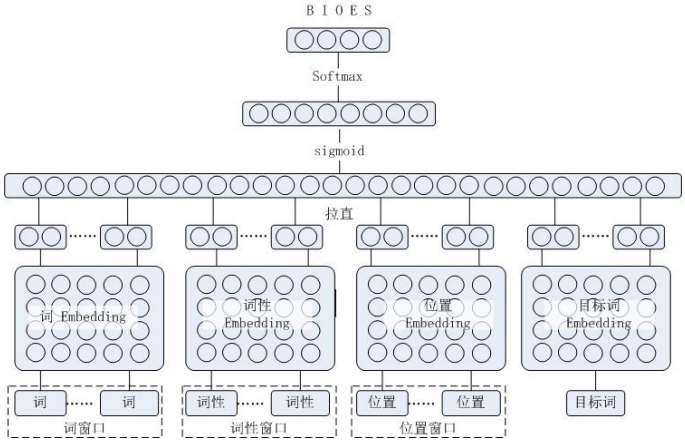

图 1 汉语框架语义识别所用的神经网络模型结构

本节提出的神经网络结构允许我们使用贪心特征窗口选择方法[15]或正交特征选择方法[2]来为每种特征选择最优的特征窗口。但由于本文主要目的是探讨 Dropout 正则化对于汉语框架语义角色识别性能的影响。因此，我们将各种特征实数向量维度和的窗口固定如下：

表 1： 模型所用特征的维度及窗口大小

特征类型	维度	窗口
词	100	[-2,2]
词性	10	[-2,2]
位置	10	[-2,2]
目标词	10	[0,0]

在表 1 中，对于词语特征，窗口[-2,2]表示当前词和前 2 个词，后 2 个词作为特征；其余特征类推；[0,0]表示仅使用当前目标词作为特征。由于语料中涉及到的目标词较少，因此，我们仅将目标词的向量维度设置为 10。

3．2 Dropout 正则化训练方法

在图 1 所给的神经网络结构，需要估计的参数主要包括：词、词性、位置和目标词的 Embedding 矩阵以及 Sigmoid 函数的参数矩阵及偏置系数；研究者通常使用人工标注语料上的模型 log 似然函数作为目标函数，然后，通过随机梯度下降的方法对上述的待估参数进行更新，直到目标函数收敛。

然而，汉语框架语义角色标注任务的可用语料规模有限。实验语料仅含 6000 多句标注句子。而且，在本文的实验中，我们仅使用一半的句子进行训练。因此，参与模型训练的句子仅有 3000 多句。训练句子的数量不足以充分地估计出模型的待估参数。这会导致最后得到的模型严重过拟合训练数据。

避免过拟合的常用方法是将正则化技术加入到训练过程中。目前，常用的正则化技术主要包括 L1, L2 和 Dropout 正则化。L1 和 L2 正则化主要是将待估参数的先验分布约束为拉普拉斯分布和正态分布。在原有的条件随机场等模型中，常用 L1 和 L2 正则化来避免过拟合。

Dropout 正则化是针对神经网络所提出来的一种正则化技术[16]。不同于 L1 和 L2 正则化，在训练神经网络时，Dropout 正则化是将神经网络结构中的隐层节点按照给定的概率进行丢弃。因为丢弃过程随机忽略隐层节点，且在每次的训练过程中每次随机忽略的隐层节点都不同，这就使每次训练的网络都是不一样的。因此，每次训练都可以看作使用一个“新”的模型；此外，隐含节点都是以一定概率随机出现，因此不能保证每个隐含节点每次都同时出现，这样权值的更新不再依赖于有固定关系隐含节点的共同作用，阻止了某些特征仅仅在其它特定特征下才有效果的情况。

本文中我们仅对 Sigmoid 层中的连接进行随机丢弃，余下的网络结构保持不变。在第 5 部分，我们对比了加入 Dropout 正则化的模型性能与不进行正则化的模型性能。实验结果表明，加入 Dropout 正则化的汉语框架语义角色识别模型的性能有着显著的提升，F 值达到 69.02%。

4 实验数据及设置

本文的实验语料来自于山西大学开发的汉语框架网络例句库。本实验语料中主要包括 25 个框架的 6692 条句子。实验语料的具体分布信息与文献[2]中的相同。为了评价本文所提模型的性能，我们采用组块 3×2 交叉验证进行实验。具体做法是，将语料库切分成 4 个大小相同的子集，然后，通过两两组合，形成组块 3×2 份交叉验证实验。组块 3×2 交叉验证在模型估计和选择的优良性能已经得到证明，具体可参考 Wang 等的工作[17]。

对于汉语框架语义角色识别模型，本文采用如下四种评价指标:

标记准确率 = 标记正确的个数/总的标记个数

准确率 = 模型识别正确的语义角色块数/模型识别出的语义角色总块数

召回率 = 模型识别正确的语义角色块数/测试集中的原有语义角色总块数

F 值 = 准确率*召回率* 2 /（准确率+召回率)

对于组块 3×2 交叉验证，我们使用 6 组实验的标记准确率，准确率，召回率和 F 值的平均值作为模型的最终指标。

在本文的神经网络结构中，我们使用[-0.5,0.5]之间的均匀分布来初始化词、词性、目标词和位置的 Embedding 矩阵；词特征的分布式表示向量为 100，其余特征的分布表示向量为 10。神经网络中除 Embedding 矩阵外的其余参数的初始值为 0；学习率设置为 0.03；隐层节点个数为 100。

5 实验结果

我们在表 2 中对比了不加 Dropout 正则化的模型性能与加入 Dropout 正则化的模型性能。当加入 Dropout 正则化后，我们分别设置了多种连接的丢弃概率。具体的实验结果见表 2。

表 2：加入 DropOut 正则化后的语义角色识别性能

Dropout 概率参数	标记准确率	准确率	召回率	F 值
0（不加 dropout）	60.98%	69.25%	59.76%	62.32%
0.4	63.46%	77.70%	53.27%	63.19%
0.5	67.54%	79.29%	56.57%	66.03%
0.6	68.75%	81.45%	59.97%	**69.07%**
0.7	68.94%	79.12%	60.10%	68.03%
0.8	68.28%	78.91%	60.15%	68.25%

从表 2 中可以看出，不加 Dropout 正则化技术的模型 F 值仅为 62.32%。然而，当加入 Dropout 增加化后，若将连接的丢弃概率设置为 0.6, 最终的模型 F 值达到 69.07%。模型的 F 值提高了近 7%。当丢弃概率加大后，模型的性能将会下降。进一步分析，我们发现，Dropout 正则化大幅度提高了语义角色块识别的准确率，对于召回率影响不大。这说明，Dropout 正则化可以有效的避免模型的过拟合现象。

上述实验中，我们使用的 Embedding 的初始值是完全随机的。为了验证 Embedding 的初始值对于模型性能的影响，我们分别使用 C&W[6], SGNS[12]以及 RNN-LM[18]训练产生的词语 Embedding 矩阵。在训练 Embedding 矩阵时，我们使用搜狗中文语料，并使用中科院自动分词工具。我们仅将词语特征的 Embedding 矩阵替换为这些算法产生的 Embedding 矩阵；余下三种特征的 Embedding 矩阵仍然保持随机。Embedding 矩阵的不同初始值得到的模型性能如表 3 所示。

表 3：不同词 Embedding 下的语义角色识别性能

词分布表征	标记准确率	准确率	召回率	F 值
随机	68.75%	81.45%	59..97%	69.07%
C & W	67.77%	79.64%	61.19%	**69.20%**
SGNS	68.31%	79.47%	60.96%	68.99%
RNN-LM	68.30%	80.07%	60.39%	68.82%

从表 3 中可以看出，加入 C&W 训练出来的词语 Embedding 矩阵后，模型的性能有了轻微上升，F 值的提升仅为 0.13%。而 SGNS 算法以及 RNN-LM 算法所产生的词语 Embedding 矩阵所对应的模型性能反而下降。我们分析，这可能是因为自动的分词信息使词语 Embedding 矩阵质量得不到保证，无法有效地体现出词语之间的语义关系。

影响语义角色识别模型性能的另一个重要的因素是神经网络训练所用的学习率。较小的学习率会导致神经网络收敛的速度变慢，但可能会使神经网络收敛到更为优良的局部最优值。表 4 中给出了不同的学习率的条件下，汉语框架语义角色识别模型的性能。

表 4：不同学习率下的语义角色标注性能

学习率	标记准确率	准确率	召回率	F 值
0.03	67.77%	79.64%	61.19%	69.20%
0.001	70.03%	84.23%	60.86%	**70.54%**
0.003	69.59%	83.36%	60.91%	70.41%
0.01	68.97%	81.70%	60.81%	69.54%
0.1	67.36%	79.46%	60.07%	68.33%
0.3	67.02%	79.32%	60.74%	68.21%

从表 4 中可以看出，减少学习率的大小可以明显提高模型的识别性能。将学习率从 0.03 调整至 0.001，模型的 F 值从 69.20%提升至 70.54%，提高 1.34%。对比准确率和召回率可以发现，减少学习率主要带来了准确率识别的提升。

表 5：与已有识别模型的结果对比

模型	标记准确率	准确率	召回率	F 值
CRF 识别模型[3]	--	73.18%	64.43%	68.52%
本论文	70.03%	84.23%	60.86%	**70.54%**

我们将本文所得到的汉语框架语义角色识别模型的最优性能与原有的汉语框架语义角色识别模型的最优性能进行比较，得到表 5 所示结果。从表中可以看出，加入特征的分布式

表示信息，并采用 Dropout 正则化来进行训练，模型的识别性能有了 2%的提升（F 值）。其中，尽管召回率有所下降，但准确率却提升了近 10%。需要指出的是，核方法是构建高性能语义边界识别的重要方法[19]，但目前并未有相关学者使用核方法来展开汉语框架语义角色识别的相关研究，因此，我们并未与该方法的性能进行对比。

6 总结及展望

本文初次尝试了使用神经网络来构建汉语框架语义角色识别模型。本文给出了一种融合多种特征信息的神经网络结构。该网络结构保留了不同特征窗口选择的灵活性。为了有效地缓解可用语料过少所产生的模型过拟合现象，本文将 Dropout 正则化的技术引入到模型的训练过程中。实验结果表明，Dropout 正则化的加入，可以有效地提升汉语框架语义角色识别模型的性能。本文进一步对词特征的 Embedding 初始值以及模型训练的学习率进行了调优。最后所得到汉语框架语义角色识别模型的 F 值达到了 70.54%，比之间的最优模型性能提高近 2%。由于通用的汉语句法分析器目前还不成熟，本文并未考虑句法层面的常用特征，例如，句法类型标记，句法子范畴特征等。

在本文中，我们并未对特征窗口选择方法进行深入探讨，并未对比 Dropout 正则化方法与 L1 及 L2 正则化方法及基于核方法的汉语语义角色标注系统性能。这些都是我们未来的研究方向。

参考文献

[1] Fillmore C J, Baker C F. Frame semantics for text understanding[C]//Proceedings of WordNet and Other Lexical Resources Workshop, NAACL. 2001.

[2] 李济洪, 王瑞波, 王蔚林, 等. 汉语框架语义角色的自动标注[J]. 软件学报, 2010, 21(4): 597-611.

[3] 宋毅君, 王瑞波, 李济洪, 等. 基于条件随机场的汉语框架语义角色自动标注[J]. 中文信息学报, 2014, 28(3): 36-47.

[4] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

[5] Pradhan S, Ward W, Hacioglu K, et al. Shallow Semantic Parsing using Support Vector Machines[C]//HLT-NAACL. 2004: 233-240.

[6] 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.

[7] Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural networks[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2015.

[8] Hong S, Noh H, Han B. Decoupled deep neural network for semi-supervised semantic segmentation[C]//Advances in Neural Information Processing Systems. 2015: 1495-1503.

[9] Shi L, Mihalcea R. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing[M]//Computational linguistics and intelligent text processing. Springer Berlin Heidelberg, 2005: 100-111.

[10] 邵艳秋, 穗志方, 吴云芳. 基于词汇语义特征的中文语义角色标注研究[J]. 中文信息学报, 2009, 23(6): 3-11.

[11] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010: 384-394.

[12] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[13] Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]//EMNLP. 2014, 14: 1532-1543.
[14] 李国臣, 党帅兵, 王瑞波, 等. 基于字的分布表征的汉语基本块识别[J]. 中文信息学报, 2014, 28(6): 18-25.
[15] 李国臣, 王瑞波, 李济洪. 基于条件随机场模型的汉语功能块自动标注[J]. 计算机研究与发展, 2010, 47(2): 336-343.
[16] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[17] Yu W, Ruibo W, Huichen J, et al. Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms[J]. Neural computation, 2014, 26(1): 208-235.
[18] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//INTERSPEECH. 2010, 2: 3.
[19] Che W, Zhang M, Aw A, et al. Using a Hybrid Convolution Tree Kernel for Semantic Role Labeling[J]. ACM Transactions on Asian Language Information Processing, 2008, 7(4).

王瑞波 (1985－),男,博士研究生,主要研究领域为统计自然语言处理.

王蔚林(1964－),男,教授,博导,主要研究领域为统计自然语言处理，统计机器学习.

李国臣(1963－),男,教授,主要研究领域为中文信息处理.

杨耀文(1990－),男,硕士研究生,主要研究领域为中文信息处理.

作者联系方式：王瑞波　山西省太原市坞城路 92 号山西大学理科楼计算中心 B508 室 030006　手机:15235127138 E-mail: wangruibo@sxu.edu.cn