基于BiLSTM聚合模型的汉语框架语义角色识别

曹学飞
山西大学
自动化与软件学院
caoxuefei@sxu.edu.cn

李济洪　王瑞波
山西大学
现代教育技术学院
{lijh; wangruibo}@sxu.edu.cn

牛倩
山西大学
自动化与软件学院
niuqian@sxu.edu.cn

摘要

目前，基于神经网络的汉语框架语义角色识别模型的性能依然较低，考虑到神经网络模型的性能受到超参数的影响，本文将超参数调优和模型预测性能的提升统一到基于BiLSTM的聚合模型框架下解决。使用正则化交叉验证进行实验，通过正则化条件约束训练集和验证集的分布差异，避免分布不一致带来的性能波动。将交叉验证得到的结果进行众数投票，以投票后的结果对不同的超参数配置进行评估，并选择若干种没有显著差异的超参数配置构成最优的超参数配置集合。然后将最优的超参数配置集合对应的子模型进行聚合，构造汉语框架语义角色识别的聚合模型。实验结果显示，本文方法的性能较基准模型显著提升了9.56%。

关键词： 框架语义角色识别；正则化交叉验证；聚合模型

Chinese Frame Semantic Role Identification Based on BiLSTM Aggregation Model

Cao Xuefei
School of Automation and Software Engineering, Shanxi University
caoxuefei@sxu.edu.cn

Li Jihong　Wang Ruibo
School of Modern Education Technology, Shanxi University
{lijh; wangruibo}@sxu.edu.cn

Niu Qian
School of Automation and Software Engineering, Shanxi University
niuqian@sxu.edu.cn

Abstract

The performance of Chinese frame semantic role identification model based on neural network is still low. Considering that the performance of neural network model is affected by hyperparameters, this paper unifies the improvement of performance and hyperparameter tuning of neural network to an aggregation model. Experiment is carried out by using regularized $m\times2$ cross-validation, and the distribution difference between the training set and the validation set is constrained by regularization conditions to avoid performance fluctuations caused by distribution inconsistency. The results obtained by cross-validation are voted, and different hyperparameter configurations are evaluated with the results of majority voting, and several hyperparameter configurations without significant differences are selected to form the optimal set of hyperparameter configurations. Then, the submodels corresponding to the optimal hyperparameter configuration set construct an aggregation model for Chinese framenet semantic role identification. Experimental results show that the performance of the

基金项目：国家自然科学基金(61806115, 62076156)

proposed method is significantly improved by 9.56% compared with the benchmark model.

Keywords: Frame semantic role identification , Regularized cross-validation , Aggregation model

1 引言

语义角色标注是语义分析中的关键技术，通过标注句子中的各种语义角色，有助于计算机准确提取句子中蕴含的语义依赖关系，从而为机器翻译、信息检索等系统提供语义支持。

汉语框架语义知识库(Chinese FrameNet，简称CFN) (刘开瑛, 2011)是以框架语义学 (Fillmore, 1976)为理论基础，以FrameNet (Baker et al., 1998)为重要参照，以真实汉语语料为事实依据构建的一个汉语词汇语义知识库，是进行汉语语义分析的一种重要资源。

框架语义学认为，词语的语义是由该词语激起的框架所描述的场景来表示的，场景中的各种参与者就是该词语所支配的语义角色。例如，对于句子“农民购买优质农用物资”，词语“购买”可以激起“商品交易”这一框架，支配两个语义角色，其中“农民”担任“购买”的“买者”角色，“优质农用物资”担任“购买”的“商品”角色。框架语义学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描述词语之间的语义关系 (王瑞波等, 2017)。进而，句子乃至篇章的语义就可以由词语激起的框架、框架之间的关系以及框架与句子中的语义角色之间的关系来表达，这就为文本的理解提供了丰富的形式化的语义信息 (宋毅君等, 2014)。

框架语义学在特定的框架下理解词语，其语义角色的描述变的更加细化和丰富，因此，CFN中的语义角色相对较多，例如，“陈述”框架就有14个语义角色。在进行汉语框架语义角色标注时，需要标注的语义角色的种类和数量都远远大于其它知识库下的标注，导致了汉语框架语义角色自动标注的难度更大。但是，丰富的语义角色可以提供更加丰富的语义信息，有助于提升计算机正确处理信息的能力。因此，构建高精度的汉语框架语义角色标注模型就成为面向汉语框架语义分析中的一个关键环节。

目前，汉语框架语义角色标注模型的精度还较低，李济洪等 (2010)经过深入分析，认为汉语框架语义角色标注的难点在于语义角色的自动识别。因此，在后续的工作中，研究者们将汉语框架语义角色标注分为语义角色识别和语义角色分类两步进行，并主要针对语义角色识别展开研究。

随着深度学习的兴起，神经网络模型也广泛应用于汉语框架语义角色识别。王瑞波等 (2017)基于词、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构建汉语框架语义角色识别模型，并采用Dropout (Nitish et al., 2014)技术来改进模型的训练过程，最终得到了70.54%的F值。在同样的语料上，党帅兵 (2015)抽取了词特征、词性特征、位置特征、目标词特征、相邻词的组合特征、相邻词性的组合特征、基本块特征，以及词、词性和位置三者之间的两两搭配特征等多种词层面特征，在基于深层神经网络的汉语框架识别模型上得到了72.89%的F值。曹学飞等 (2022)构建了基于BiLSTM (Graves and Schmidhuber, 2005)的深度神经网络模型，使用词特征、词性特征、目标词特征和目标词的位置特征，采用3×2交叉验证 (Wang et al., 2014)进行实验，得益于BiLSTM优良的表示学习能力以及对特征的优化设计 (Cao et al., 2019)，该模型得到的F值达到了77.72%，显著高于之前其他工作的结果。

然而，针对汉语框架语义角色识别这一任务，上述基于神经网络的研究也面临以下两个问题。

- 神经网络模型的性能依赖于超参数的配置 (Reimers and Gurevych, 2017)，如何选取好的超参数使得模型性能达到最优并且稳健，即超参数调优是汉语框架语义角色识别的一个关键问题。传统的超参数调优通常是将数据集切分为训练集、验证集和测试集，将每一种不同的超参数配置看作一个独立的模型，在训练集上训练，在测试集上进行测试，检验某一种超参数配置下的模型的性能评价指标是否提高。但是，数据集的随机切分方式可能导致训练集和测试集的分布差异较大，使得模型的预测性能并不稳定，常常得到不可靠的超参数比较的结论。

- 由于计算资源的制约，早期的相关工作都是在CFN的例句库的一个子集(包含25个框架、6692条标注例句，以下简称小语料)上展开，近年来，为了和之前的方法进行对

比，研究者们延续了对这一小语料的使用。而CFN目前已标注的例句库包含了约4万条例句(以下简称大语料)，涉及205个框架，在大语料上，汉语框架语义角色识别的F值最高仅为65.31% (曹学飞等, 2022)，性能仍然较低，这是由于在框架语义学中，不同框架下的语义角色的种类和数量都不同，大语料中包含了205个框架，远大于小语料中的25个框架，这导致了框架语义角色识别的难度更大。因此，在CFN的大语料上提升框架语义角色识别的性能是目前需要解决的另一个重要问题。

基于以上分析，本文在CFN的大语料上进行汉语框架语义角色识别研究，并将上述两个问题统一到一个基于BiLSTM的聚合模型的框架下解决。由于在CFN的大语料上，针对汉语框架语义角色识别任务，曹学飞等 (2022)的工作得到了目前最优的性能，本文将该方法作为实验对比的基准方法，因此，选择了该方法中采用的BiLSTM神经网络模型来构建汉语框架语义角色识别的聚合模型。①对于第一个问题，即超参数调优。与传统的调优方法选出一个最优超参数配置不同，本文提出了应该选择若干个性能上没有显著差异的超参数配置构成“最优的超参数配置集合”。具体来讲，首先基于正则化交叉验证(Regularized $m\times2$ Cross Validation，简记为$m\times2$ RCV)进行实验，$m\times2$ RCV是一种带约束切分的m次2折交叉验证的模型训练与验证方法，它通过分布差异度量函数来均衡训练集、验证集的分布差异，避免训练集、验证集分布不一致带来的性能波动。对每一种不同的超参数配置，$m\times2$ RCV可以训练得到$2m$个模型(简记为子模型)，并可得到全部语料中每一条标注例句的m个预测序列，对m个预测序列进行众数投票，以投票结果作为最终的预测序列，并和真实的标记序列比较，从而评估不同超参数配置的性能。然后，将所有的超参数配置按照评估性能从高到低排序，再次进行增量式的投票，目的是选择“最优的超参数配置集合”，即对每一条标注例句，依次将排序后的h个超参数配置对应的$h\times m$个预测序列再次进行众数投票，检验投票后得到的性能是否进一步提升，如果性能提升，则h递增1继续以上操作，如果不再提升，则此时h个超参数配置构成了“最优的超参数配置集合”。②对于上文提到的第二个问题，即在大语料上汉语框架语义角色识别性能的提升，本文方法充分利用了每一种超参数配置在调优阶段得到的子模型，将“最优的超参数配置集合”中的h个超参数配置对应的$h\times2m$个子模型分别在公共测试集上进行测试，然后采用众数投票对$h\times2m$个结果进行聚合，从而构造汉语框架语义角色识别的投票聚合模型(见图1)。实验结果表明，本文提出的方法能在实现超参数调优的同时，还能通过对多个子模型的高效聚合，提升模型的预测性能和稳健性。

本文的组织结构如下：第2节描述了汉语框架语义角色识别任务，并介绍了本文实验使用的BiLSTM神经网络模型；第3节详细描述了本文提出的聚合模型；第4部分介绍了实验的相关设置；第5部分给出了实验结果及分析；最后总结了全文，并给出了下一步的研究方向。

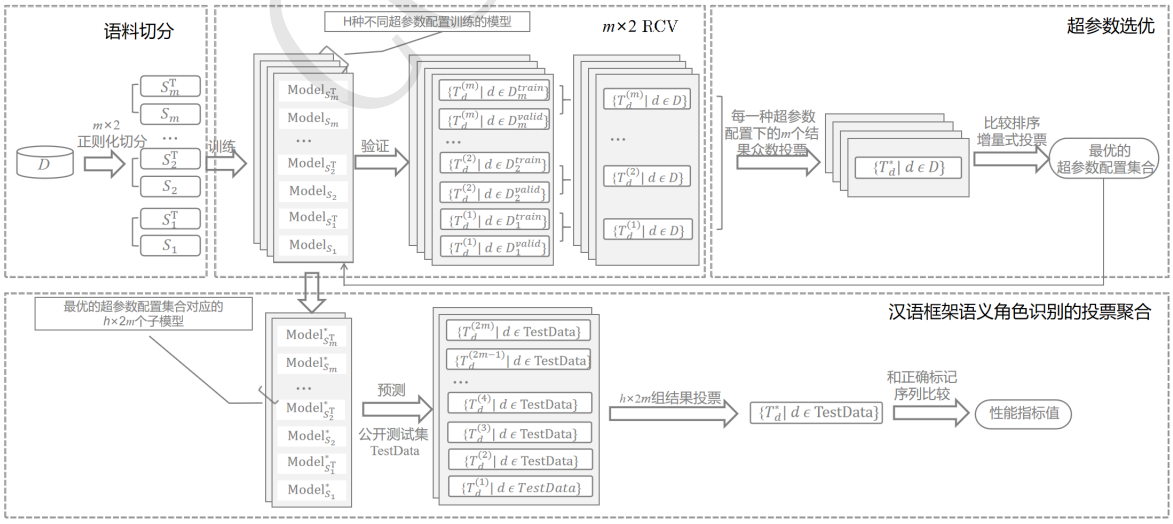

图 1：汉语框架语义角色识别的聚合模型

2 汉语框架语义角色识别任务

汉语框架语义分析包括框架识别和框架语义角色标注两部分，其中，框架识别指识别出句子中能够激起框架的成分(目标词)，并自动标注其所属的框架; 框架语义角色标注又可分为角色识别和角色标注两个步骤，即首先识别句中的哪些成分可以构成目标词所支配的语义角色，然后再对识别出的语义角色进行类别标注 (李济洪等, 2010)。

2.1 角色识别任务的形式化描述

将一条汉语句子看作是一个以词为单位的长度为N的序列，记为$S = w_1, w_2, \cdots, w_N$，$w_i$表示句子中的第$i$个词，给定句子中的目标词$w_t$，对句中每个词标记一个合适的标签$t_i$，$t_i \in \{\text{B, I, O}\}$表示句子$S$中第$i$个词对应的语义角色的边界标签，其中，B标签表示对应的词是一个语义角色块的开始词，I标签表示对应的词是一个语义角色块的中间词或结尾词，O标签表示对应的词不属于任何一个语义角色块。这样可以得到一个标签序列$T = t_1, t_2, \cdots, t_N$，基于$T$可以重构出句子$S$中的语义角色块，从而将语义角色识别转化为一个序列优化问题：$T^* = \arg\max_T \mathrm{P}(T = t_1, t_2, \cdots, t_N)$，针对该优化问题，本文构造了一个基于BiLSTM的神经网络模型进行求解(见图2)。

2.2 基于BiLSTM的语义角色识别模型

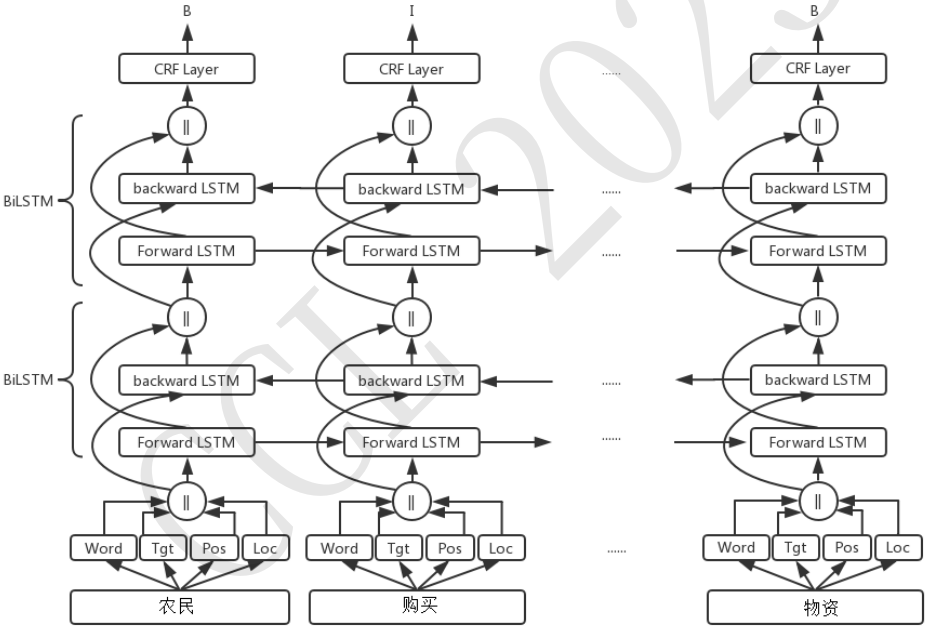

图 2: 基于BiLSTM的汉语框架语义角色识别模型

图2所示的模型包含三部分：输入层，BiLSTM层和CRF层。在输入时，将一条汉语句子看做以词为基本单位的一个序列送入模型，经过BiLSTM层的学习，在CRF层输出一个带有B、I、O标签的序列。基于输出的标签序列重构句子中的语义角色块进而完成语义角色的自动识别。

3 汉语框架语义角色识别的聚合模型

本文提出的基于BiLSTM的汉语框架语义角色识别的聚合模型如图1所示，包括以下四个模块。

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

3.1 语料切分

超参数选优，本质上是对不同超参数配置下的模型性能的比较，而$m\times2$交叉验证方法是常用的模型比较方法 (Wang et al., 2017; 王瑞波等, 2019)。$m\times2$交叉验证是指将数据集做m次随机切分，实施m次2折交叉验证。研究发现，基于随机切分的交叉验证进行模型比较，所得出的结果并不可靠 (Bergkirkpatrick et al., 2013; Rodríguez et al., 2010)。特别是，针对汉语框架语义角色识别任务，李济洪等 (2010)通过对CFN语料的分析发现：不同框架所支配的语义角色的种类和数量是不同的，如果对CFN语料随机切分，可能会导致框架、语义角色的分布不同。例如，包含了较多语义角色的标注例句切分在训练集中，而含有较少语义角色的标注例句切分在验证集中，或者某些框架对应的标注例句全部切分到训练集(验证集)中，这显然会增大性能评价指标的方差，在模型比较时产生不可靠的结论。

为解决语料的随机切分可能带来的性能波动，本文在$m\times2$交叉验证的基础上设计正则化条件约束训练集、验证集的分布差异，使得切分得到的训练集、验证集的分布较为均衡。具体来讲：假定语料D由$|D|$条标注例句组成，即$D=\{d_1,d_2,\cdots,d_{|D|}\}$，$D$上某次对半切分记为$\{D^{train},D^{valid}\}$，其中，$D^{train}\cup D^{valid}=D$，$D^{train}\cap D^{valid}=\emptyset$；$m\times2$交叉验证的切分集可记为：$P=\langle S_i,S_i^T\rangle$，其中，$S_i=(D_i^{train},D_i^{valid})$，$S_i^T=(D_i^{valid},D_i^{train})$，$i=1,2,\cdots,m$。$\langle S_i,S_i^T\rangle$为一个切分对，$S_i^T$为$S_i$的对折切分，某个切分下的$D_i^{train}$被称为训练集，$D_i^{valid}$为验证集。对于汉语框架语义角色识别任务，本文引入两个正则化切分条件：①框架在D_i^{train}和D_i^{valid}之间的差异要尽可能一致；②任意两个不同的D_i^{train}和$D_i^{valid}(i\neq j)$之间重叠的标注例句要尽量少且一致。

对于第一个正则化条件，本文使用离散随机变量分布一致性检验的卡方统计量来度量训练集和验证集之间框架的分布差异，设随机变量F表示框架名，其取值为离散集合$\{f_1,f_2,\cdots,f_J\}$，则框架在训练集和验证集上的分布差异，可用如下卡方统计量来度量：

$$\chi^2=\sum_{j=1}^{J}\frac{n^{train}(r_j^{train}-r_j)^2+n^{valid}(r_j^{valid}-r_j)^2}{r_j},\tag{1}$$

式(1)中，n^{train}和n^{valid}为其在训练集及验证集上出现的频次，r_j、r_j^{train}和r_j^{valid}为第j种取值f_j在数据集、训练集及验证集上的频率。实验中以单位自由度的差异度量函数χ^2/J来作为训练集和验证集上框架的分布差异度量指标，并经验性的选择指标值小于等于1的切分。对于第二个正则化条件，王瑞波等 (2019)已证明了任意两个不同的D_i^{train}和$D_j^{train}(i\neq j)$之间重叠的标注例句约为总例句数的1/4时，可以减小$m\times2$交叉验证的方差，本文采用王瑞波等 (2019)给出的方法来满足这一正则化条件。经过上述正则化处理，可以使得训练集与验证集中在框架分布上均衡，得到更为稳健的实验结果。

3.2 $m\times2$ RCV实验

在$m\times2$交叉验证的基础上通过正则化条件约束训练集、验证集的分布差异，称之为正则化交叉验证(简称$m\times2$ RCV) (王瑞波等, 2019)。

对$m\times2$ RCV的某个切分对$\langle S_i,S_i^T\rangle$，可以基于S_i的训练集D_i^{train}训练得到模型Model_{S_i}，然后对S_i的验证集D_i^{valid}中的例句进行预测，得到预测序列集合$T_d^{(i)}$，其中$d\in D_i^{valid}$。由于S_i^T为S_i的对折切分，将训练集和验证集交换进行实验，可基于D_i^{valid}训练得到模型$\text{Model}_{S_i^T}$，进而得到D_i^{train}中所有例句的预测的标记序列集合。因此，对每一种不同的超参数配置，$m\times2$ RCV可以训练得到$2m$个子模型，同时得到全部语料（训练集和验证集）中每条标注例句的m个预测的标记序列：$T_d^{(i)}$, $i=1,2,\cdots,m$, $d\in D$。

3.3 超参数调优

采用众数投票分别对语料中所有标注例句的m个预测序列进行投票聚合(以预测标记为单位投票)，假定对某条例句中的第n个词的m个预测标记的投票结果为t_n^*，投票准则如下：

$$t_n^*=l_{\arg\max_j\sum_{i=1}^{m}\mathbb{I}(t_n^i=l_j)},\tag{2}$$

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

其中$i=1,2,\cdots,m$，l_j为标记集合里的第j个标记，$\mathbb{I}(\cdot)$为指示函数。使用投票得到的的标记序列和该例句的真实标注序列进行比较，评估某种超参数配置下的模型的预测性能，进而可以比较得到不同超参数配置的优劣。

实际上，若干种不同的超参数配置对应的模型，其预测性能可能并没有统计意义上的显著差异。因此，本文提出了可以选择h个没有显著差异的超参数配置构成“最优的超参数配置集合”。对于h的取值，本文也给出了一种简单有效的方法：将所有的超参数配置按照评估结果从高到低排序后再次进行增量式的众数投票，即根据排序结果，依次将h个超参数配置对应的$h\times m$个预测序列进行投票，如果投票后性能进一步提升，则h递增1继续以上操作，直到性能不再提升，则此时的h个超参数配置构成了“最优的超参数配置集合”。

3.4 子模型聚合

在$m\times 2$ RCV下，超参数调优阶段每一种超参数配置均会训练得到$2m$个子模型，“最优的超参数配置集合”中的h种超参数配置可以得到$h\times 2m$个子模型，对这些子模型进行聚合构造汉语框架语义角色识别的聚合模型，即将这些子模型分别在CFN的公共测试集上进行测试，然后对$h\times 2m$个结果采用众数投票，将投票结果作为聚合模型的预测结果。

4 实验设置

4.1 语料

本文选用山西大学开发的汉语框架语义知识库(CFN)的例句库作为实验语料。该例句库包含了约4万条标注好的汉语句子，并且提供了按照约8:1:1比例切分的训练集(31526条例句)、验证集(3947条例句)和测试集(4022条例句)。针对汉语框架语义角色识别，曹学飞等 (2022)在这一语料上得到了目前最优的性能，本文将该方法作为实验对比方法，也同样采用上述切分的测试集评估本文提出的聚合模型性能。实验时，将CFN例句库提供的训练集和验证集合并，然后按照第3节的描述进行$m\times 2$ RCV进行实验(m取15)。

4.2 模型的特征及超参数设置

使用BiLSTM神经网络进行汉语框架语义角色识别时，通常可以设置一些特征丰富例句中词的信息。曹学飞等 (2022)对句子中的每个词添加了4个候选特征，包括当前词特征、当前词的词性特征、句子中的目标词特征和当前词相对目标词的位置特征(当前词在目标词左边或右边)，将该四个候选特征连同另外2个BiLSTM模型的设计选项(BiLSTM的层数和是否添加CRF层)统一看做需要调优的超参数。本文沿用了同样的设置，详细说明见表1。

表 1: 特征及超参数设置

特征	取值	说明
词	R_100, G_100	分别用随机的100维向量或GloVe (Pennington et al., 2014)预训练的100维向量来表示词
目标词	10，20	分别用随机的10维或随机的20维向量来表示目标词
词性	-，20	不使用当前词的词性特征，或用随机的20维向量来表示当前词的词性
位置信息	-，10	不使用当前词的位置特征，或用随机的10维向量来表示位置信息
BiLSTM层数	1，2	模型只采用1层BiLSTM 网络或采用2层堆叠的BiLSTM
CRF	0，1	0表示模型顶层不添加CRF层，1表示模型添加CRF层

4.3 实验的其它设置

图2所示的BiLSTM模型采用了和曹学飞等 (2022)同样的设置，包括：使用随机梯度下降算法进行训练，进行了100次的迭代，每次使用10条例句对参数进行更新(即batch-size为10)，初始学习率为0.015，学习率衰减系数为0.05；dropout rate为0.5，BiLSTM层的节点数为200。

此外，考虑到$m\times 2$ RCV对每种超参数配置都需要进行$2m$次实验，如果对表1中所有可能的超参数配置进行完全实验，所需的计算量较大，例如，本文实验中m为15，则完全实验次数为$15\times 2\times 2^6$。因此，为减少实验次数，提高超参数选优效率，本文采用$L_8(2^7)$正交表从所有可

表 2: $L_8(2^7)$正交表设计的8种超参数组合

实验号	列号						
	1(词)	2(目标词)	3(词性)	4(位置信息)	5(BiLSTM层数)	6(CRF)	7
1	R_100	10	20	-	2	1	-
2	R_100	10	20	10	1	0	-
3	R_100	20	-	-	2	0	-
4	R_100	20	-	10	1	1	-
5	G_100	10	-	-	1	1	-
6	G_100	10	-	10	2	0	-
7	G_100	20	20	-	1	0	-
8	G_100	20	20	10	2	1	-

能的超参数配置中选择相关性较低的超参数配置进行调优，即从2^6种中选出了8组有代表性的超参数配置来安排实验，实验次数减少为15×2×8。正交表的设计见表2（每一个实验号所在的行表示一种不同的超参数配置）。

4.4 性能评价指标

对于汉语框架语义角色识别任务，通常采用按识别语义角色块的P、R和F值作为模型性能的评价指标，定义如下：

P=模型识别正确的语义角色块个数(TP)/模型识别出的语义角色块总个数(TP+FP)，

R=模型识别正确的语义角色块个数(TP)/原有语义角色块总个数(TP+FN)，

F=2×P×R/(P+R)。

5 实验结果及分析

我们首先按照传统调优的方式，调优选出一种最优超参数配置(5.1节)，然后将该超参数配置对应的子模型进行聚合，初步说明聚合模型的优势(5.2节)。进一步我们选出了多个超参数配置构成“最优的超参数配置集合”(5.3节)，最后基于该集合中的多个超参数配置对应的子模型构造聚合模型(5.4节)。

5.1 最优超参数配置

对语料中的每一条例句，15×2 RCV可以得到15个预测序列，表3中的“均值”(标准差)表示某一种超参数配置下，对15×2 RCV的15组预测序列分别进行评估得到的均值和标准差(F值)，“投票结果”表示将每条例句对应的15个预测的标记序列以标记为单位进行众数投票，得到该例句投票后的标记序列，然后对语料中所有例句的投票后的标记序列进行评估的结果。

表 3: 不同超参数配置下的预测结果(%)

实验号	均值	标准差	投票结果	投票−均值
1	69.46	0.45	74.65	5.19
2	62.86	0.61	68.39	5.53
3	63.31	0.77	69.73	6.42
4	61.31	0.70	68.06	6.75
5	62.90	0.58	67.42	4.52
6	63.83	0.85	69.69	5.86
7	62.38	0.40	67.71	5.33
8	70.75	0.43	**74.97**	4.22

表3结果显示，实验号8对应的超参数配置在15×2 RCV下投票得到了最高的F值，在传统的超参数调优中，记其为最优的超参数配置。表3的最后一列显示，8组不同超参数配置上的投票

结果相比均值都有显著的提升，最低提升了4.22%，平均提升了5.48%，这也可以说明投票聚合方法在汉语框架语义角色识别上的有效性。

5.2 基于最优超参数配置的聚合模型

5.2.1 聚合模型的性能

本节讨论基于5.1节给出的最优超参数配置下的聚合模型，将8号超参数配置(最优超参数配置)对应的15×2个子模型分别在CFN的测试集上进行测试，再对15×2个结果进行众数投票，以最终的投票结果度量该聚合模型在汉语框架语义角色识别任务上的性能。实验结果如图3和表4所示。图3的结果显示，聚合模型的预测结果达到了74.87%(F值)，优于任意一个子模型单独测试得到的结果(均值为70.39%)。

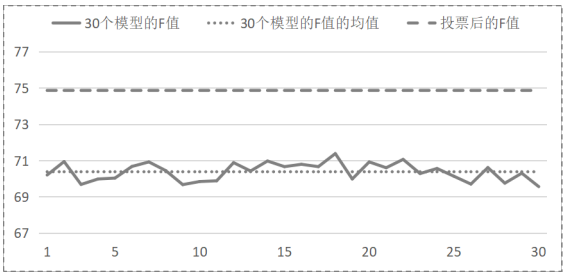

图 3: 基于最优超参数配置的聚合模型的性能

表 4: 基于投票的框架语义角色识别结果

	P	R	F
均值	71.38	69.46	70.39
聚合	77.10	72.76	**74.87**
曹学飞等 (2022)	64.16	66.49	65.31

表4的结果显示，与曹学飞等 (2022)的结果相比，本文的方法大幅提升了9.56%。与15×2个子模型测试结果的均值相比，聚合模型得到的P和R也都有明显的提高。进一步分析可知，R的提高得益于聚合模型可以得到更多的“识别正确的语义角色块”(TP)，见图4。P的提高是由于聚合模型在提高“识别正确的语义角色块”(TP)个数的同时，可以减少“识别错误的语义角色块”(FP)的数目，见图5。

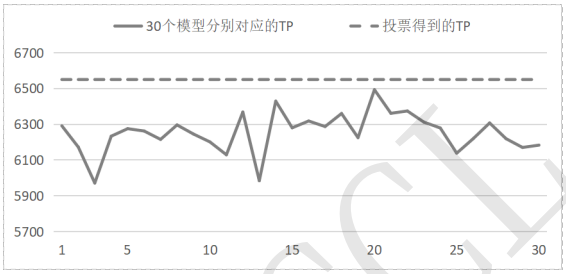

图 4: 模型投票前后TP的对比

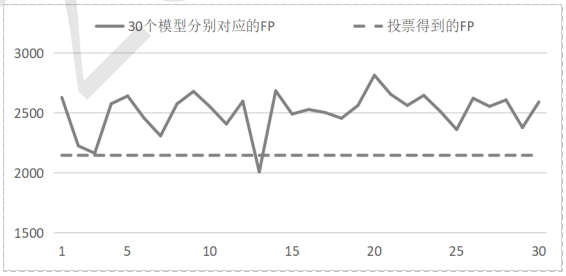

图 5: 模型投票前后FP的对比

5.2.2 显著性检验

Wang and Li (2019)给出了一种适用于m×2交叉验证下计算P、R和F值的分布函数及置信区间的方法，并说明了可以从两个结果的置信区间有无交叉来判别差异的显著性，当两个结果在1-α置信区间上没有交叉时，那么二者在置信水平α下有显著差异。本文利用该方法分别计算了聚合模型和曹学飞等 (2022)的P、R以及F值的置信区间。表5为显著性水平α=0.05下P、R和F值的置信区间的结果对比，显然，聚合模型获得的性能提升相比曹学飞等 (2022)的结果是显著的。

表 5: 置信区间的对比

	聚合	曹学飞等 (2022)
P	[76.19, 77.98]	[63.16, 65.14]
R	[71.83, 73.67]	[63.49, 65.47]
F	[74.14, 75.57]	[63.50, 65.13]

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

5.2.3 聚合模型的稳健性

本节我们从置信区间宽度和多次重复实验两个方面探讨聚合模型的稳健性。

(1)置信区间宽度的比较

我们分别计算了最优超参数配置(8号超参数配置)对应的15×2个子模型在测试集上预测得到的F值的置信区间，图6为所有子模型以及聚合模型对应的置信区间宽度值排序后的折线图，显然，聚合模型(投票)的置信区间宽度小于任一个参与投票的子模型的置信区间宽度，而置信区间的宽度越小，意味着对应模型的稳健性越好。

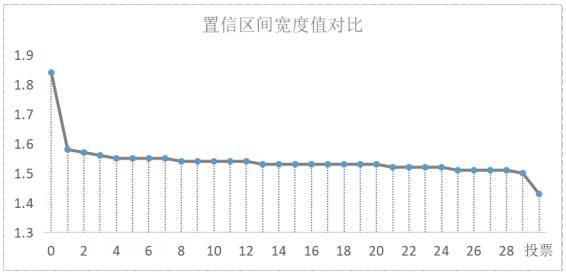

图 6: 置信区间宽度值对比

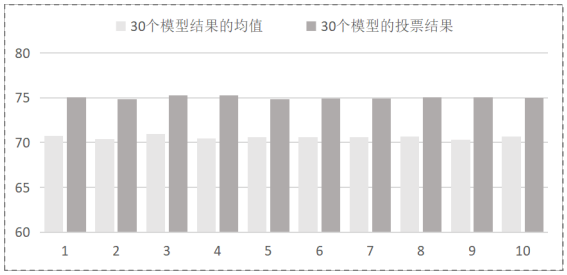

图 7: 10次重复实验结果

(2)多次重复实验的结果比较

为了进一步验证聚合模型的稳健性，本文按照如下方式在CFN的语料集进行了10折交叉验证，通过10次重复实验的结果来比较说明：

- 将CFN语料切分10份，每次选取其中9份组合，然后按照本文方法在9份组合的语料上进行15×2 RCV，然后投票选择一种最优超参数配置
- 对每次调优得到的一种最优超参数配置对应的15×2个子模型进行聚合，即将它们分别在剩下的1份语料上测试，对15×2个测试结果进行投票用来评估最终性能。

如图7所示，本文的方法可以得到非常稳健的输出结果，10次重复实验得到的聚合模型的结果(F值)，均显著优于其对应的15×2个子模型预测结果的均值，且10次聚合结果的标准差仅为0.15%。

5.3 最优的超参数配置集合

本文基于m×2 RCV进行实验，而m的选择一定程度上影响了聚合模型最终的性能，如图8所示，随着m的增大(见图8的横轴坐标)，聚合模型的预测性能在一个较大的增幅后($m < 7$)，逐渐进入一个平缓的增加态势。如果继续增大m($m > 15$)，即使能够继续保持缓慢增加的态势，也可能获得较小的性能增益，但无疑会大大增加超参数调优的时间，带来更大的计算开销。

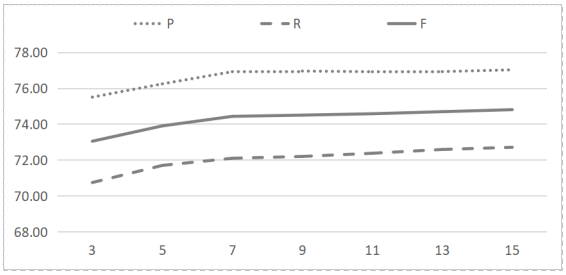

图 8: m的不同取值下投票聚合得到的结果

表 6: 构造“最优的超参数配置集合”时的投票结果

	P	R	F
8	76.97	73.07	74.97
8 & 1	77.48	73.05	**75.20**
8 & 1 & 3	76.92	72.29	74.53

直观上来讲，m×2 RCV保证了任意两次切分得到的训练集之间的重叠例句(样本)的数目尽可能少且一致，这也就意味着参与聚合的子模型来自不同的训练样本，拥有不同的预测能力。m的增大又意味着参与投票的子模型的数量增加，从而提升了聚合模型的性能。我们可以换个思路，是否可以不增大m但又可以增加参与投票的子模型的数量。考虑到若干种不同的超

参数配置对应的模型，其预测性能可能并没有统计意义上的显著差异。因此，本文进一步提出了在超参数调优时，可以选择多个没有显著差异的超参数配置构成“最优的超参数配置集合”，而不仅仅是选择一种最优超参数配置。

以表3的结果来说明“最优的超参数配置集合”的选择，8种超参数配置的优先排序如下“8 > 1 > 3 > 6 > 2 > 4 > 7 > 5”，“最优的超参数配置集合”的初始值即为8号超参数配置。然后，按照优先顺序进行增量式的投票聚合，即将在15×2 RCV时得到的1号超参数配置对应的预测序列和8号超参数配置对应的预测序列一起进行投票，再和正确标记序列对比，从表6的结果可知，本次增量投票可以得到了更高的F值，因此，当前“最优的超参数配置集合”应该包含8号和1号两种超参数配置。按照优先次序，继续使用3号超参数配置对应的预测序列进行增量式投票，表6的结果显示F值不再增加。这样我们可以得到最终的“最优的超参数配置集合”，即包含了8号和1号两种超参数配置。

5.4 基于最优的超参数配置集合的聚合模型

将“最优的超参数配置集合”中的每一种超参数配置在m×2 RCV时得到的子模型共同作为聚合模型的子模型，这样可以大大增加参与投票的子模型的数量。本文实验中，我们将“最优的超参数配置集合”中8号和1号超参数配置对应的子模型(2×15×2个)在公共测试集进行测试，并将预测结果进行众数投票，得到的结果如表7所示。显然，基于“最优的超参数配置集合”的聚合模型的结果(P、R和F)均比基于“最优超参数配置”的聚合模型的结果更好，虽然F值仅提高了0.23%，但二者之间有显著性差异的置信度达到了71.78%[0]。

表 7: 基于“最优的超参数配置集合”的聚合结果

编号	TP	FN	FP	P	R	F
8	6550	2452	1946	77.10	72.76	74.87
8 & 1	6575	2427	1918	**77.42**	**73.04**	**75.16**

表 8: 调优时，不同超参数配置的置信区间

8号	1号	3号
[74.86, 75.35]	[74.40, 74.90]	[69.48, 69.99]

分析发现，“最优的超参数配置集合”中的8号和1号超参数配置，在调优时分别对应的投票结果没有显著差异(即表3中实验号8和1对应的投票结果)。表8给出了显著性水平α=0.05下二者的F值的置信区间对比，结果显示，8号和1号的置信区间有交叉部分，这意味着这两种超参数配置并无显著差异，而它们又都和3号超参数配置有显著差异，因此，本文的“最优的超参数配置集合”实质上是超参数调优时得到的没有显著差异的若干个最好的超参数配置，进而可以将它们在m×2 RCV时训练得到的子模型聚合，通过增加性能上没有显著差异的子模型的数量提升聚合模型的性能。

6 总结

本文提出了一个基于BiLSTM的汉语框架语义角色识别的聚合模型，首先，采用m×2 RCV进行试验，可以尽可能降低实验结果对语料切分的敏感性，进而对m组结果进行众数投票，以投票结果来评估不同超参数配置的优劣。然后根据超参数配置的优先次序继续进行增量式投票，得到“最优的超参数配置集合”。最后利用“最优的超参数配置集合”对应的所有子模型构建汉语框架语义角色识别的聚合模型。与基准方法相比，该聚合模型可以显著提升汉语框架语义角色的识别性能，说明了该聚合模型的有效性。此外，本文创新性的提出了超参数调优时应该选择多种性能上没有显著差异的超参数配置构成“最优的超参数配置集合”，这样可以增加参与聚合的子模型的数量，从而提升聚合模型的性能。本文使用基于BiLSTM的神经网络模型作为子模型进行聚合，一方面是受计算平台性能的限制，另一方面是为了和基于BiLSTM的基准方法更好的对比。理论上，任何一种神经网络模型都可以作为该聚合模型框架中的子模型，因而，利用BERT等预训练模型作为子模型构建汉语框架语义角色识别的聚合模型也是我们下一步的研究计划。

参考文献

刘开瑛. 2011. 汉语框架语义网构建及其技术研究. 中文信息学报, 25(6):46–52.

[0]置信度的计算采用了 (Wang and Li, 2019)中的贝叶斯检验方法。

Charles J. Fillmore. 1976. *Frame semantics and the nature of language.* Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280:20–32.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. *The Berkeley FrameNet Project.* Association for Computational Linguistics, 86–90.

王瑞波，李济洪，李国臣，杨耀文. 2017. 基于*Dropout*正则化的汉语框架语义角色识别. 中文信息学报, 31(1):147–154.

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

李济洪, 王瑞波, 王蔚林. 2010. 汉语框架语义的自动标注. 中文信息学报, 21(4):597–611.

Srivastava Nitish, Hinton Geoffffrey, Krizhevsky Alex, Sutskever Ilya, and Salakhutdinov Ruslan. 2014. *Dropout: a simple way to prevent neural networks from over-fifitting. Journal of Machine Learning Research*, 15(1):1929–1958.

党帅兵. 2015. 基于词分布表征的汉语框架语义角色识别研究. 山西大学.

曹学飞, 李济洪, 王瑞波, 牛倩, 王钰. 2022. 基于稳健设计的双向长短期记忆神经网络模型的调优方法. 应用概率统计, 38(3):317–332.

Alex Graves and Jurgen Schmidhuber. 2005. *Framewise phoneme classifification with bidirectional LSTM and other neural network architectures.* Neural Networks, 18:602–610.

Yu Wang, Jihong Li Ruibo Wang and Huichen Jia. 2014. *Blocked 3×2 cross-validated t-Test for comparing supervised classifification learning algorithms.* Neural Computation, 26(1):208–235.

Xuefei Cao, Jihong Li, Ruibo Wang, Yu Wang, Qian Niu and Junfeng Shi. 2019. *Calibrating GloVe model on the principle of Zipf's law.* Pattern Recognition Letters, 125(7):715–720.

N Reimers and I Gurevych . 2017. *Reporting score distributions makes a diffffference: performance study of LSTM-networks for sequence tagging.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 338–348.

Ruibo Wang, Yu Wang, Jihong Li, Xingli Yang and Jing Yang. 2017. *Block-Regularized m × 2 Cross-Validated Estimator of the Generalization Error.* Neural Computation, 29(2):519–554.

王瑞波, 王钰, 李济洪. 2019. 面向文本数据的正则化交叉验证方法. 中文信息学报, 33(5):54–65.

Bergkirkpatrick T, Burkett D and Klein D. 2013. *An Empirical Investigation of Statistical Significance in NLP.* IEEE Geoscience & Remote Sensing Letters, 13(3):457–461.

Rodríguez J D, Perez A, Lozano J A. 2010. *Sensitivity analysis of k-fold cross validation in prediction error estimation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3):569—575.

Pennington J, Socher R and Manning C. 2014. *GloVe: global vectors for word representation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.

Ruibo Wang and Jihong Li. 2019. *Bayes Test of Precision, Recall, and F1 Measure for Comparison of Two Natural Language Processing Models.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4135–4145.