基于神经耦合模型的异构词法数据转化和融合

黄德朋　李正华†　龚晨　张民

苏州大学计算机科学与技术学院，苏州 215006; † 通信作者, E-mail: zhli13@suda.edu.cn

摘要　有监督机器学习方法需要一定规模的人工标注数据来学习模型参数。一方面，人工标注数据的规模通常会显著影响模型的性能，另一方面，人工标注数据的代价非常大。因此，如何充分利用已有的异构人工标注数据一直是学术界的一个关注方向。Li et al. (2015)提出了一种基于传统离散特征的耦合序列标注模型，可以有效对异构词法数据进行转化和融合。将其方法扩展到基于 BiLSTM 的深度学习框架上，直接在两个异构训练数据上训练参数，测试阶段则同时预测两个标签序列。在词性标注、分词词性联合标注两个任务上进行了大量实验。结果表明，与多任务学习方法和传统耦合模型相比，神经耦合模型在利用词法异构数据方面更优越。神经耦合模型在异构数据转化和融合两个场景上都取得了更高的性能。

关键词　耦合模型；BILSTM；深度学习；词性标注；分词

中图分类号

Neural Network Coupled Model for Conversion and Exploitation of Heterogeneous Lexical Annotations

HUANG Depeng,　LI Zhenghua†,　GONG Chen,　ZHANG Min

School of Computer Science and Technology, Soochow University, Suzhou 215006;

† Corresponding author, E-mail: zhli13@suda.edu.cn

Abstract Supervised machine learning methods require a certain amount of manual labeling to learn model parameters. On the one hand, the scale of manual annotations often significantly affects the performance of the model. On the other hand, the cost of manual annotations is very high. Therefore, how to make full use of existing heterogeneous manual data has always been a focus of the academic community. Li et al. (2015) propose a coupled sequence labeling model based on traditional discrete features, which can effectively convert and exploit heterogeneous lexical data. The authors extend their approach under the BiLSTM-based deep learning framework. The neural coupled model learn its parameters directly on two heterogeneous training data, and predict two optimal sequences simultaneously during the test phase. The authors have conducted a lot of experiments on the part-of-speech(POS) tagging task and the joint word segmentation and POS(WS&POS) tagging task. The results show that neural coupled approach is superior to other methods for exploiting heterogeneous lexical data, including the multi-task learning method and the traditional discrete-feature coupled model. Neural coupled model achieves higher performance on both scenarios, i.e., annotation conversion and boost the final target-side tagging accuracy by exploiting heterogeneous data.

Key words　Coupled model; BILSTM; deep learning; part-of-speech tagging; word segmentation

词法分析任务是一个根据给定句子，给出其词边界、词性以及实体的过程。作为中文信息处理的最基础任务，词法分析的效果直接影响到上层任务（如句法分析[1]、信息抽取[2]等）的性能。近年来，深度学习在自然语言处理各个任务上都取得了很好的效果。在词法分析任务上，如分词和词性标注[3]，基于 BiLSTM-CRF 的序列标注方法和基于传统离散特征的序列标注方法相比，由于其强大的句子表征能力，显著提高了分析性能，已经成为新的主流方法。

除了深度学习模型带来的积极影响，在有监督学习场景中，数据的质量和规模也显著影响着词法分析的性能。由于汉语中存在很多遵守不同标注规范的数据，于是相比于费时费力地重新构建人工标

国家自然科学基金(61525205, 61876116, 61702518)资助

收稿日期：0000-00-00；修回日期：0000-00-00；网络出版日期：0000-00-00

注数据，如何充分利用异构数据来提升统计模型的性能成为了研究者们非常关注的问题。目前常用于词法分析的异构数据有 CTB、PKU（PD）、MSR，本文采用由 Xue et al. (2005)[4]构造的宾州中文树库（Penn Chinese Treebank，CTB）和由北大计算语言学研究所构造的人民日报语料库（People's Daily，PD）[5]进行实验，数据规范间的词性以及分词差异如表 1 所示。

表 1 异构数据举例

Table1 A example of heterogeneous annotations

语料	例句									
CTB-词性	特别是/AD			我/PN	国/NN	经济/NN		下滑/VV		。/PU
PD-词性	特别/d		是/v	我国/n		经济/n		下滑/v		。/w
CTB-分词词性	特 B@AD	别 I@AD	是 E@AD	我 S@PN	国 S@NN	经 B@NN	济 E@NN	下 B@VV	滑 E@VV	。 S@PU
PD-分词词性	特 B@d	别 E@d	是 S@v	我 B@n	国 E@n	经 B@n	济 E@n	下 B@v	滑 E@v	。 S@w

为了充分利用异构数据提升统计模型性能，前人提出了几种较为有效的方法，如 Jiang et al. (2009)[6]提出的指导特征方法，Qiu et al. (2013) [7]提出的线性耦合模型以及 Chen et al. (2016)[8]提出的基于深度学习的多任务学习方法。尤其是，Li et al. (2015)[9]首次尝试并提出了基于传统离散特征的耦合序列标注方法，Li et al. (2016)[10]在其基础上进一步优化，在不同规模的异构数据上进行了大量实验分析，不仅在词法分析上取得了显著的性能提升，在异构词法数据的转化和融合两个子任务上也有更佳的表现。

考虑到深度学习方法在词法分析任务上带来的积极影响，本文首次尝试并提出了基于深度学习的神经耦合标注方法。在词法分析任务以及异构数据的转化融合任务上的实验表明：1）和基于深度学习的单语料基准模型相比，神经耦合模型更加有效；2）和多任务学习方法相比，耦合标注模型能够在取得较高词法分析性能的同时，拥有更强的异构数据转化融合能力；3）相比于基于传统离散特征的 CRF 耦合模型，基于深度学习的神经耦合模型可以取得更高的词法分析和异构数据转化融合性能。

1 基于双向长短期记忆网络的标注模型

依据目前词法分析基于深度学习上的主流方法，我们将长短期记忆（Bidirectional Long Short-Term Memory, BiLSTM）网络作为基准标注模型。如图 1 所示，该模型包含表示层、编码层、得分层和预测层四部分。表示层（Input Layer）将数据中的词或字处理成特征向量作为模型的输入，编码层通过 BiLSTM 网络获取句子中每个词或字的上下文特征，得分层通过多层感知器（Multi-layer Perception, MLP）计算每个标签的得分，预测层使用 Softmax 函数预测每个词或字的标签。

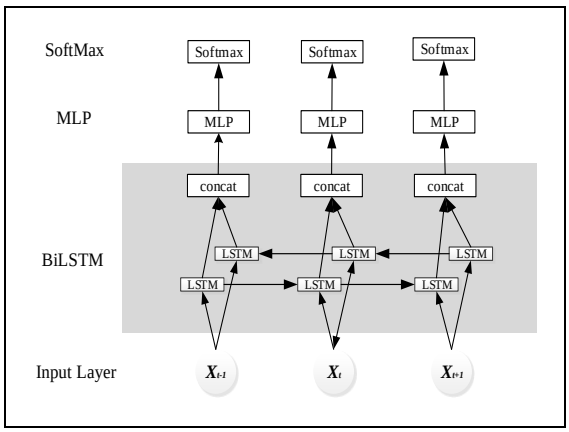

图 1 长短期记忆网络模型

Fig.1 BiLSTM model

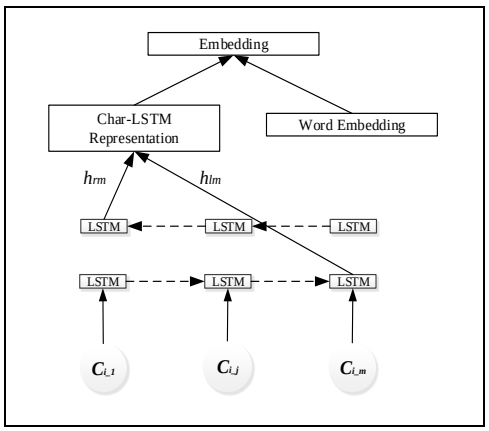

图 2 词的准确表示

Fig.2 Representation of word

1.1 表示层

在词性标注任务上，我们在表示层同时获取词向量和字向量特征并经过一定的编码后作为网络的输入。词向量采用预训练的方式，而字向量随机初始化并进行单独的编码处理。给定一个句子 $S=\{w_1,w_2,w_3,\ldots,w_n\}$，$w_i$ 表示句子中的第 i 个词，n 表示句子中词的个数，对于每个词 $w_i=\{c_{i\_1},c_{i\_2},c_{i\_3},\ldots c_{i\_m}\}$，$c_{i\_j}$ 表示词 w_i 中的第 j 个字，m 表示该词中字的个数。如图 2 所示，我们将 w_i 中所有字的随机初始化向量输入到 BiLSTM 中，并将 BiLSTM 中前向后向各自的最后一个隐藏层输出$\boldsymbol{h}_{lm}$和$\boldsymbol{h}_{rm}$拼

接到 w_i 对应的词向量$\boldsymbol{e}_w$后面，得到 w_i 的最终特征向量$\boldsymbol{X}_i$，那么$\boldsymbol{X}_i$可以表示为：

$$\boldsymbol{X}_i = \boldsymbol{e}_w \oplus \boldsymbol{h}_{lm} \oplus \boldsymbol{h}_{rm}, \quad (1)$$

而在分词词性联合标注任务上，我们随机初始化当前字的 unigram 向量与 bigram 向量。给定一个以字为单位的句子 $S=\{c_1,c_2,c_3,\ldots,c_n\}$，$c_i$ 表示句子中的第 i 个字，n 表示句子中字的个数。对于每个字 c_i，我们窗口化前一个字 c_{i-1} 与当前字，以获取当前字的局部特征向量 bigram，并与当前字的 unigram 向量拼接，那么 c_i 的最终特征向量$\boldsymbol{X}_i$可以表示为：

$$\boldsymbol{X}_i = \boldsymbol{e}_{c_{i-1}c_i} \oplus \boldsymbol{e}_{c_i}, \quad (2)$$

其中，$\boldsymbol{e}_{c_{i-1}c_i}$表示当前字的 bigram 特征向量，$\boldsymbol{e}_{c_i}$表示当前字的 unigram 特征向量。

1.2 编码层

编码层使用 LSTM 对句子信息进行编码。我们将句中词或字的最终特征表示向量$\boldsymbol{X}_i$作为 LSTM 的输入,对整个句子序列进行编码得到当前词或字的全局信息$\boldsymbol{h}_i$，形式上可以定义为：

$$\boldsymbol{i}_i = \sigma(\boldsymbol{W}_{in} \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{X}_i] + \boldsymbol{b}_{in}), \quad (3)$$

$$\boldsymbol{f}_i = \sigma(\boldsymbol{W}_{fg} \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{X}_i] + \boldsymbol{b}_{fg}), \quad (4)$$

$$\boldsymbol{o}_i = \sigma(\boldsymbol{W}_{out} \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{X}_i] + \boldsymbol{b}_{out}), \quad (5)$$

$$\boldsymbol{c}_i = \boldsymbol{f}_i \cdot \boldsymbol{c}_{i-1} + \boldsymbol{i}_i \cdot \tanh(\boldsymbol{W}_c \cdot [\boldsymbol{h}_{i-1}, \boldsymbol{X}_i] + \boldsymbol{b}_c), \quad (6)$$

$$\boldsymbol{h}_i = \boldsymbol{o}_i \cdot tanh(\boldsymbol{c}_i), \quad (7)$$

其中，$\boldsymbol{i}_i$，$\boldsymbol{f}_i$，$\boldsymbol{o}_i$，$\boldsymbol{c}_i$分别表示第 i 个词对应的输入门、遗忘门、输出门和细胞状态输出，$\boldsymbol{X}_i$和$\boldsymbol{h}_i$表示第 i 个词或字对应的输入向量和隐藏层输出。σ和tanh为激活函数，$\boldsymbol{W}$和$\boldsymbol{b}$分别对应各个门的权重以及偏置。

LSTM 的隐藏状态只是从过去获取信息，而从不考虑未来的信息。为了能够编码两个方向的句子信息，我们将前向后向两个 LSTM 的隐层输出拼接在一起，得到词 w_i 的 BiLSTM 隐藏状态表示$\boldsymbol{h}_i$：

$$\boldsymbol{h}_i = \overrightarrow{\boldsymbol{h}_i} \oplus \overleftarrow{\boldsymbol{h}_i}, \quad (8)$$

其中，$\overrightarrow{\boldsymbol{h}_i}$和$\overleftarrow{\boldsymbol{h}_i}$分别表示 BiLSTM 隐藏层的前向、后向输出。

1.3 得分层

得分层使用 MLP 来计算每个标签的得分，将 BiLSTM 的输出$\boldsymbol{h}_i$作为 MLP 的输入，从而得到句子中每个词或字对应的所有标签的得分$\boldsymbol{Score}_i$：

$$\boldsymbol{Score}_i = \boldsymbol{W}_{mlp} \cdot \boldsymbol{h}_i \oplus b_{mlp}, \quad (9)$$

其中，$\boldsymbol{W}_{mlp}$和b_{mlp}分别表示 MLP 层的权重和偏置。

1.4 预测层

我们尝试并对比了 CRF 和 local loss，发现性能差别不大，且在耦合模型的实验上，将 CRF 作为预测层对速度的影响比较大。为了平衡速度与性能，我们在所有模型上都采用 Softmax 函数预测最终标签并采用交叉熵（CrossEntropy）函数作为目标函数。

对于给定词 w_i，共有 m 种可能词性，那么每个标签的得分经过 Softmax 的归一化处理后的输出为：

$$y_j' = softmax(score_j) = \frac{e^{score_j}}{\sum_{k=1}^{m} e^{score_k}}, \quad (10)$$

其中y_j'表示第 j 个标签经过归一化处理之后的可能概率。则目标函数为：

$$loss = -\sum_{j=1}^{m} y_j \cdot log\, y_j', \quad (11)$$

其中y_j表示期望的概率分布。

2 基于多任务学习的异构序列标注数据利用

为了能够有效融合异构数据之间的差异，并比较各个模型利用异构数据所能达到的性能，我们在基准模型上实现了 Chen et al. (2016)[8]提出的基于神经网络多任务学习的异构数据模型。

多任务学习方法相比于基准模型的优势在于能够同时利用几个异构数据混合训练模型参数，能够共享表示层和编码层的参数，并独立训练各自的 MLP 层参数，从而预测各自的标签，在扩大了训练语料的基础上提升了基准模型的性能。

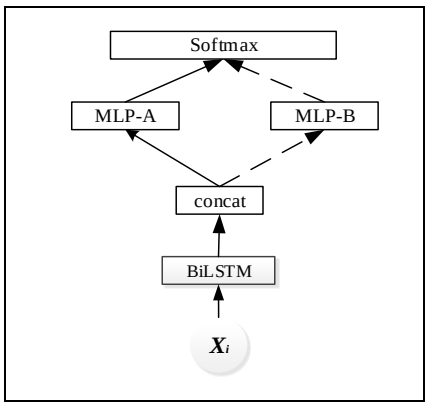

图 3 多任务学习模型
Fig.3 Multi-Task Learning model

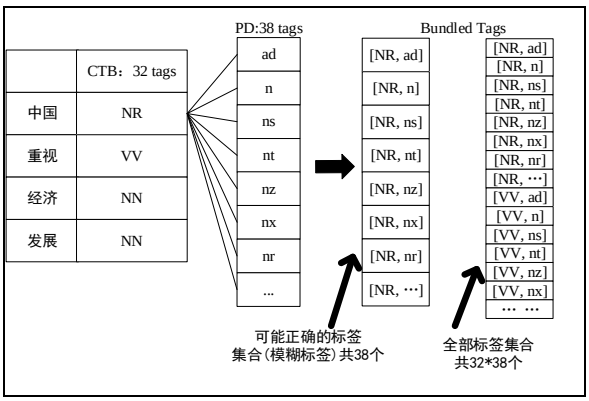

图 4 词性模糊标签示例
Fig.4 A example of POS ambiguous labeling

多任务学习方法与基准模型的网络框架基本一致，区别在于经过 BiLSTM 编码之后，会根据语料来源信息选择对应的 MLP 层计算标签得分。如图 3 所示，在两个异构数据的多任务学习模型上，共有两个独立的 MLP 层，当训练语句来自 A 语料时，BiLSTM 的输出会输入给 MLP-A 计算标签得分，当训练语句来自 B 语料时，BiLSTM 的输出将输入到 MLP-B 中计算标签得分。模型相对较为简单，当输入只有单一语料时，模型本身相当于基准模型。

3 基于神经网络的耦合标注方法

由于多任务学习模型单独预测各个数据上的标签序列，只有输入层和 BiLSTM 层同时由两个数据训练得到，MLP 层单独训练，所以与 Li et al. (2015)[9]提出的耦合模型相比，异构数据训练的参数较少。

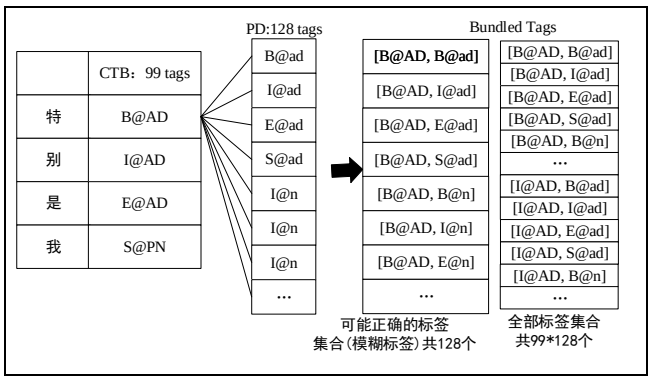

图 5 分词词性联合模糊标签示例
Fig.5 A example of WS&POS ambiguous labeling

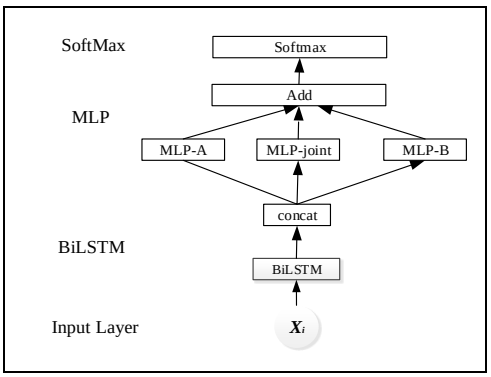

图 6 神经耦合模型
Fig.6 Neural coupled model

Li et al. (2015)[9]提出的基于传统离散特征的耦合序列标注模型的具体做法是将两个异构数据的标签进行捆绑（例如词性标签“[NN,n]”），并在扩大的捆绑标签集合上进行建模。如图 4、图 5 所示，通过考虑所有的捆绑标签可能性将单一标签映射转化为一组捆绑标签（称之为模糊标签）。我们将其扩展到基于 BiLSTM 的神经网络模型上，并根据模糊标签重新设计适用的目标函数。

如图 6 所示，与基准模型和多任务学习模型不同的是，得到 BiLSTM 的输出后同时输入到三个 MLP 中，分别为计算两个规范上独立标签得分的 MLP-A、MLP-B 以及一个计算捆绑标签得分的 MLP-joint。最终根据耦合映射关系将两个独立标签的得分与联合标签得分相加，得到输入给目标函数的得分，句子 S 中第 i 个词的联合词性标签为$[t^a, t^b]$的得分为：

$$Score(s, i, [t^a, t^b]) = Score_{joint}(s, i, [t^a, t^b]) + Score_{sep\_a}(s, i, t^a) + Score_{sep\_b}(s, i, t^b), \quad (12)$$

其中，$Score_{joint}(s, i, [t^a, t^b])$表示句子 S 中第 i 个词被标为联合标签$[t^a, t^b]$的得分，$Score_{sep\_a}(s, i, t^a)$、$Score_{sep_b}(s, i, t^b)$分别表示句子 S 中第 i 个词被标为 A 语料标签集合中独立标签t^a、B 语料标签集合中独立标签t^b的得分。

相应地，由于每个词的正确词性不止一个，所以本文对目标函数进行了扩展。例如 CTB5 的词性标签数量为$|T_a|$=32，PD 的词性标签数量为$|T_b|$=38，所以 CTB5 语料中的每个词的正确模糊标签有 38

个，而 PD 语料中的每个词的正确模糊标签有 32 个。于是目标函数（Cross Entropy）相应的扩展为：

$$loss = -log \sum_{j=1}^{k} y_j, \quad (13)$$

其中，k 表示正确的模糊标签的数量，y_j为正确模糊标签的得分经过 Softmax 归一化之后的概率。

4 实验

4.1 实验设置

在这一节，我们在词性标注任务和分词词性联合标注任务上分别建立了大量的实验去验证我们方法的有效性。

数据集设置 我们采用了三个包含分词信息和词性信息的数据集来训练并评估构建的词性标注模型和分词词性联合标注模型，分别为 CTB5 和两个不同规模的 PD 语料，规模如表 2 所示。两个 PD 语料分别由 Li et al. (2016)[10]和 Li et al. (2015)[9]提供，本文简称 PD1、PD2。由于缺少 Chen et al. (2016)[8]的数据，所以多任务学习也在这两组异构数据上进行实验。此外我们还采用了 Li et al. (2016)[10]提供的 1000 句来自 PD2 的人工标注数据“newly labeled”来测试模型的异构数据转化能力。这 1000 句人工标注数据中，共有 27942 个词标有 PD 规范的词性标签，而只有 5769 个词人工标有 CTB 规范的标签。

表 2 数据规模统计

Table 2 Data size statistics

#corpus		#sentences	#tokens	
CTB	train	16,091	437,911	
	dev	803	20,454	
	test	1,910	50,319	
PD1	train	46,815	1,097,839	
	dev	2,000	46,182	
	test	5,000	118,174	
PD2	train	273,883	6,499,208	
	dev	1,000	23,427	
	test	2,500	58,301	
newly labeled		1,000	with CTB tags	with PD tags
			5,769	27,942

评价指标 实验中词性实验的评价指标为词性准确率（Accuracy），而分词词性联合的评价指标为：准确率（Precision，P）、召回率（Recall，R）、综合指标值（F-measure，F）。其中分词词性联合的评价不仅仅是根据分词的 B、I、E、S 标签计算 P、R、F 值，还需要词中每个字的词性标签一致。模型转化能力评价采用 Li et al. (2016)[10]的做法，即用训练出的最佳模型预测出 1000 句人工标注数据中每个词在 CTB 规范上的词性，并在其中人工标注 CTB 规范词性的 5769 个词上评价词性准确率。

模型设置 我们基于 PyTorch 框架实现多个方法，并使用 Adam 更新算法对模型进行更新[11]。为了合理地混合两个异构数据，我们采用 Li et al. (2014)[12]提出的 corpus weighting 作为数据混合训练方法，每次迭代分别从两份训练数据中随机抽取 M、N 个句子混合训练，在各自的 dev 数据上词性准确率或者分词词性标注的 F 值迭代 I 次不提高就完全停止。我们根据不同比例混合 CTB5 与 PD1 的训练集进行实验对比后采用 M=N=16,091（CTB5 的全部训练数据）以及 I=10 的实验设置。模型中词向量和字向量维度均为 100，学习率设为 0.001，dropout 为 0.55。

4.2 词性标注结果

为了验证神经耦合模型的有效性，我们首先在词性标注任务上做了实验，并与前人工作进行对比。为了更好地与 Li et al. (2015)[9]和 Li et al. (2016)[10]的结果进行比较，我们在 CTB5、PD2 和 CTB5、PD1 两组异构数据集上都分别训练了三个模型，结果如表 3、表 4 所示。

表 3 不同方法在 CTB5 和 PD1 上的词性标注准确率（Accuracy）

Table 3 The precision of part-of-speech tagging on CTB5 and PD1 by different methods

模型		CTB5		PD1	
		Dev	Test	Dev	Test
Baseline BiLSTM		94.87	94.33	96.20	96.16
Multi-Task Learning		95.46	95.05	**96.27**	**96.31**
Neural Coupled Model		**95.81**	**95.45**	96.07	96.19
Li et al. (2016)	Baseline CRF	-	94.07	-	95.82
	Coupled CRF	-	94.83	-	95.90

表 4 不同方法在 CTB5 和 PD2 上的词性标注准确率（Accuracy）
Table 4 The precision of POS tagging on CTB5 and PD2 by different methods

模型		CTB5		PD2	
		Dev	Test	Dev	Test
Multi-Task Learning		95.95	95.89	97.53	97.42
Neural Coupled Model		**95.99**	**95.96**	**97.64**	**97.56**
Li et al. (2015)	Baseline CRF	94.28	94.10	-	-
	Coupled CRF	95.10	95.00	-	-

在基准词性标注的方法上，BiLSTM 模型要比基于传统离散特征的 CRF 模型略高。而在利用了异构数据之后，无论是基于多任务学习的方法，还是本文提出的神经耦合模型，都相比基准模型都有明显的提高。尤其是对数据规模相对较小的 CTB5 数据，多任务学习方法在两组异构数据实验性能上分别提升 0.72%和 1.56%，神经耦合模型能够提升 1.12%和 1.63%。结果表明异构数据的使用确实能够帮助提升词性标注的准确性。

与多任务学习方法相比，在数据规模相对较大的 PD 上，神经耦合模型的结果与之基本持平，甚至在 PD1 上还略低。而在 CTB5 的测试集上，尤其是与 PD1 组成异构数据进行的实验上，神经耦合模型比多任务学习方法要高出 0.40%。这是由于 CTB5 数据规模较小，PD 数据规模较大，在异构数据模型中，CTB5 能够学到更多的 PD 知识从而提升性能。而 PD 只能从 CTB5 中学到很少，对标注性能几乎无影响。

与基于传统离散特征的耦合 CRF 模型相比，在两组异构数据实验上，基于深度学习的耦合模型有很明显的提高。除了在 CTB5 上分别有 0.62%和 0.96%的提升，即使是在 PD1 上，也有 0.29%的提升。说明了在词性标注任务上，深度学习方法确实要比传统离散特征方法更具备优势。也更加验证了本文提出的深度学习耦合模型更能充分利用异构数据。

在 PD1 和 PD2 上，几个模型的性能相差都不明显，无论是多任务学习方法还是耦合标注方法，都不能有效利用异构数据提升性能。说明当数据规模较大，且另一份数据规模与之相比有一定的差距时，从中得到的帮助微乎其微。

4.3 分词词性联合标注结果

为了进一步地验证模型的有效性，在词性标注任务之外，我们还在分词词性联合标注任务上做了一组对比实验。为了与 Li et al. (2016)[10]的实验结果对比，我们采用一样的数据集 CTB5、PD1，并训练了基准模型、基于多任务学习的异构数据模型和基于深度学习的耦合标注模型，结果如表 5 所示。

表 5 不同方法在 CTB5 和 PD1 上的分词词性联合标注结果(PRF)
Table 5 The result(PRF) of WS&POS tagging on CTB5 and PD1 by different methods

模型		CTB5-test			PD1-test		
		P	R	F	P	R	F
Baseline BiLSTM		88.96	89.66	89.31	**92.77**	**92.65**	**92.71**
Multi-Task Learning		90.40	90.56	90.48	92.48	92.41	92.45
Neural Coupled Model		90.66	**90.66**	90.51	92.28	92.28	92.37
Li et al. (2016)	Baseline CRF	89.60	89.38	89.49	92.74	92.20	92.47
	Coupled CRF	**90.68**	90.49	**90.58**	92.70	92.19	92.44

与词性标注结果不同的是，和传统离散特征耦合 CRF 模型相比，神经耦合模型的性能与之基本持平。和基于传统离散特征的基准 CRF 相比，基准 BiLSTM 模型的性能相当，CTB5 测试集上略低，而 PD1 测试集上略高。实验结果表明，在标签数量较大的分词词性联合任务上，基于深度学习的 BiLSTM 和基于传统离散特征的 CRF 表现基本一致，且 BiLSTM 较为容易受到数据规模的影响。

和多任务学习方法相比，神经耦合模型也没有明显的提升，基本保持持平。结果表明在分词任务上，耦合模型相对于多任务学习方法并无明显优势。这是因为虽然耦合模型在得分层训练了更多的参数，但是编码层几乎一样，导致耦合模型的优势更多在于数据的转化能力，并非表现在单独的词法分析性能上。

在 CTB5 的测试集上，无论是基于传统离散特征的耦合模型还是基于深度学习的耦合模型，又或是多任务学习方法，相比基准模型均有 1.2%左右的明显提升。表明了异构数据模型确实能够有效提升分词性能。

在 PD1 的测试集上，各个模型性能几乎一致，最简单的 BiLSTM 基准模型性能反而最佳。这主要因为 PD1 规模相比 CTB 较大，本身性能已经较高，而较小规模的 CTB 对 PD1 的帮助不大，在 CTB 和 PD1 上进行的词性标注实验与 Li et al. (2016)[10]的实验结果均表明了这一点。而由于实验受到随机初始化以及 batch 大小等一些因素的影响，各模型性能会有一定的轻微浮动。

4.4 模型的异构数据词性转化

除了在基本的词性标注任务和分词词性联合标注任务上验证了神经耦合模型的有效性，我们还在模型对异构数据的词性转化上进行了实验分析。结果如表 6 所示。

表 6 不同方法的模型转化准确率

Table 6 Model conversion accuracy of different methods

模型		PD-to-CTB 转化准确率
Baseline BiLSTM		91.49
Multi-Task learning (with PD1)		93.10
Neural Coupled model (with PD1)		93.36
Multi-Task learning (with PD2)		94.16
Neural Coupled model (with PD2)		**94.54**
Li et al. (2015)	Baseline CRF	90.59
	Coupled CRF (with PD2)	93.90

基于深度学习的基准 BiLSTM 模型和基于传统离散特征的基准 CRF 模型均由 CTB5 训练而得，结果表明，与在词性标注任务上的对比结果类似，基于深度学习方法的基准模型在词性转化能力上要明显优于传统离散特征方法，取得了 0.90%的性能提升。

在利用 CTB5、PD1 这组异构数据训练的模型上，神经耦合标注模型与多任务学习方法相比于基准模型都有很大的提升，分别提升 1.61%、1.87%。在利用 CTB5、PD2 这组异构数据训练的模型上，不仅是多任务学习方法的转化性能进一步提升，基于深度学习和基于传统离散特征的异构数据模型相比于基准模型的提升也都更为明显。表明了异构数据模型能够非常有效地提升对数据的转化融合能力。

与多任务学习方法相比，无论是在 CTB5、PD1 数据上进行的实验，还是在更大规模的 CTB5、PD2 数据上，结果都表明由于神经耦合模型更全面的考虑了异构数据之间不同规范标签集合之间的联系，从而具有更强的词性转化性能。且在较大规模的异构数据上，神经耦合模型相比多任务学习方法有 0.38%的准确率提升。

与 Li et al. (2015)[9]基于传统离散特征的耦合 CRF 模型相比，基于深度学习方法的耦合模型在词性转化上有 0.64%的提升，达到了几个模型的最佳。结果表明，基于深度学习方法的耦合模型能够更佳的完成异构数据的词性转化与融合。

4.5 实验结果分析

相同模型在不同数据上的差异表明，在异构的两组数据上，规模更大的 CTB5、PD2 能够取得更高的性能提升。而在异构数据内部，规模较大的 PD 数据并不能从 CTB5 中得到明显的有效帮助。

对比词性标注中各个模型在 CTB5 上的表现，我们发现神经耦合模型能够取得最佳的性能。表明基于深度学习的耦合模型具备比基于传统离散特征的耦合模型更强的表征能力，能够在词性标注任务上取得更好的分析性能。神经耦合模型性能优于多任务学习方法也说明，耦合模型更能够充分利用异构数据达到提升统计模型性能词性标注准确率的目的。

在异构数据的词性转化和融合子任务上，由于耦合模型更全面的考虑了异构规范信息，同时利用两组标签信息学习得分参数，所以能够取得更好的数据转化性能，从而进行更好地数据融合。又因为深度学习方法更强大的句子表征能力，所以相对于基于传统离散特征的耦合模型，神经耦合模型能够更优地表征异构数据之间的耦合关系，从而能够取得比 Li et al. (2015)[9]提出的基于传统离散特征的耦合模型更优秀的异构数据词性转化和融合能力。

5 相关工作

由于历史原因，同一种语言遵守不同的标注规范，虽然质量较高，却不能互相融合。引发了目前许多研究者致力于设计一种有效的方法来充分利用异构标注数据，尤其是在中文处理任务中。

Jiang et al. (2009)[6]首先提出了类似于堆叠学习(Nivre and McDonald, 2008)[13]的指导特征（guide-feature）方法，用于在 CTB 和 PD 数据上的分词词性联合标注任务。Sun et al. (2012)[14]进一步扩展了指导特征方法，并且提出了一种更加复杂的子词叠加的方法。与他们在一种资源上生成特征加到另外一种资源上的做法不同，Li et al. (2015)[9]提出的耦合模型能够在两个数据上都使用联合特征，利用联合捆绑标签训练模型参数，更好的融合了两个异构数据之间的共性。

Qiu et al. (2013) [7]提出了线性耦合模型，但是他们的模型只使用了单独标签特征。Li et al. (2015)[9]提出的耦合模型类似于因子 CRF(Sutton et al., 2004)[15]，在某种意义上说捆绑标签可以分解为两个连接

的潜在变量。在这项工作中，耦合序列标注模型联合处理两个具有不同标注规范的相同任务，捆绑标签可以分解为两个连接的独立标签。

目前已经有很多关于模糊标注的研究用于分类任务(Jin and Ghahramani,2003)[16]、序列标注任务(Dredze et al., 2009)[17]以及分析任务(Riezler et al., 2002[18]; Täckström et al.,2013[19])。而 Li et al. (2015)[9]的工作提供了一种利用类似模糊标注、建立捆绑标签去训练模型的方法，其中一个句子只包含单种规范的标签。Li et al. (2016)[10]在 Li et al. (2015)[9]的模型基础上，对数据的可能标签集合进行了修剪，在更小更精确的标签空间上进行建模。而我们模型扩展到了神经网络上，在未修剪的标签集合上建模，不仅在未损失效率的同时训练了更多的参数，而且利用改进的交叉熵函数使得模型更加高效。

此外，异构数据的使用也与多任务学习密切相关，而多任务学习的方法是利用共享标签的交互特征同时学习多项相关任务。Chen et al. (2016)[8]提出了基于神经网络的多任务学习模型用于词性标注，共享底层模型参数，并单独计算两个数据各自的标签得分并预测最终标签序列。而我们的神经耦合模型，所有的模型参数几乎都是共享，能够得到更好的训练。

最终的结果也表明，本文提出的基于深度学习的耦合模型相比前人的方法，能够在词法分析任务上取得持平甚至更优的性能。而在对数据的转化融合上，我们的方法具备明显的优势。

6 结语

在有监督学习场景中，训练数据的质量和规模显著影响着词法分析的性能。由于存在许多遵守不同词法规范的汉语数据，而且这种异构数据比较容易获得，于是如何充分利用异构数据来提升词法分析任务上统计模型的性能成为了受到广泛关注的研究问题。

除了受到数据规模的影响，词法分析任务的性能也受到模型选择的影响。近年来，由于深度学习在自言语言处理各个任务上均取得了良好的效果，在词法分析任务如分词和词性标注上，基于深度学习的 BiLSTM-CRF 模型已经成为了目前的主流方法。

为了充分利用异构数据达到提升模型性能的目的，前人已经提出了几种较为有效的方法。最为简单有效的方法是 Chen et al. (2016)[8]提出的基于深度学习的多任务学习方法，以及在 Li et al. (2015)[9]提出的基于传统离散特征的耦合序列标注模型。这两种方法均能在较少增加模型复杂度的基础上较大提升词法分析的性能。

本文借鉴前人的工作，将耦合模型扩展到深度学习框架上。利用 BiLSTM 模型强大的句子表征能力，在捆绑标签上进行建模，同时计算两组独立标签的得分与捆绑标签的得分，并重新设计适用于模糊标注的目标函数。我们与前人的工作进行对比，最终结果表明，本文的模型在词性标注任务上所能达到的性能更高，在分词词性联合任务上也能保持持平。另外，由于本文提出的模型计算了捆绑标签得分，更充分的融入了不同的规范信息，所以在异构数据的词性转化与融合任务上有非常明显的优势。

参考文献

[1] McDonald R, Crammer K, Pereira F. Online large-margin training of dependency parsers. ACL. 2005: 91–98.

[2] Mccallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. ICML. 2000: 591–598.

[3] Berger A L, Pietra S A D, Pietra V J D. A Maximum Entropy approach to Natural Language Processing. ACL. 1996: 39–71.

[4] Xue Naiwen, Xia Fei, Chiou Fudong, et al. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural Language Engineering. 2005:207–238.

[5] Yu Shiwen, Lu Jianming, Zhu Xuefeng, et al. Processing norms of modern chinese corpus. Technical report. 2001.

[6] Jiang Wenbin, Huang Liang, Liu Qun. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging: a case study. ACL. 2009:522–530.

[7] Qiu Xipeng, Zhao Jiayi, Huang Xuanjing. Joint Chinese word segmentation and POS tagging on heterogeneous annotated corpora with multiple task learning. EMNLP. 2013:658–668.

[8] Chen Hongshen, Zhang Yue, Liu Qun, et al. Neural network for heterogeneous annotations. EMNLP. 2016:731–741.

[9] Li Zhenghua, Chao Jiayuan, Zhang Min, et al. Coupled sequence labeling on heterogeneous annotations: POS tagging as a case study. Association for Computational Linguistics. 2015:1783–1792.

[10] Li Zhenghua, Chao Jiayuan, Zhang Min, et al. Fast coupled sequence labeling on heterogeneous annotations via context-aware pruning. EMNLP. 2016:753–762.

[11] Kingma, Diederik P, Ba J. Adam: A method for stochastic optimization. Computer Science. 2014.

[12] Li Zhenghua, Zhang Min, Chen Wenliang, et al. Ambiguity-aware ensemble training for semi-supervised dependency parsing. ACL. 2014:457–467.

[13] Nivre, Joakim, McDonald, et al. Integrating graph-based and transition-based dependency parsers. ACL. 2008:950–958.

[14] Sun Weiwei, Wan Xiaojun. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations. ACL. 2012: 232–241.

[15] Sutton C, McCallum A, Rohanimanesh K. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. Journal of Machine Learning Research. 2007, 8(5): 693–723.

[16] Jin Rong, Ghahramani Zoubin. Learning with multiple labels. NIPS. 2003: 921–928.

[17] Dredze M, Talukdar P P, Crammer K. Sequence learning from data with multiple labels. Workshop Co-Chairs. 2009:39.

[18] Riezler, Stefan, King. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. ACL. 2002: 271–278.

[19] Täckström O, McDonald R, Nivre J. Target language adaptation of discriminative transfer parsers. NAACL. 2013: 1061–1071.