

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 为了扩大人工标注数据的规模, 从而提高模型性能, 尝试充分利用已有的异构人工标注数据训练模型参数。将Li等2015年提出的耦合序列标注方法扩展到基于BiLSTM的深度学习框架, 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, 在异构数据转化和融合两个场景上都取得更高的性能。

关键词 耦合模型; BiLSTM; 深度学习; 词性标注; 分词

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 In order to expand the scale of manual annotated data and thereby improve model performance, we attempt to make full use of existing heterogeneous annotations to learn model parameters. We extend coupled sequence labeling model proposed by Li et al. (2015) under the BiLSTM-based deep learning framework. The neural coupled model learn its parameters directly on two heterogeneous training data, and predicts two optimal sequences simultaneously during the test phase. A lot of experiments have been conducted 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

词法分析任务是根据给定的句子, 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很好的效果。在词法分析任务(如分词和词性标注^[3])上, 与基于传统离散特征的序列标注方法相比, 基于BiLSTM-CRF的序列标注方法由于其强大的句子表征能力, 显著地提高了分析性能, 已成为主流方法。

除深度学习模型带来的积极影响外, 在有监督学习场景中, 数据的质量和规模也显著地影响词法分析的性能。由于汉语中存在很多遵守不同标注规范的数据, 因此相比于费时费力地重新构建人工标注数据, 如何充分利用异构数据来提升统计模型的性能受到研究者们更多的关注。目前, 常用于词法分析的异构数据有CTB, PKU(PD)和MSR。本文采用Xue等^[4]构造的宾州中文树库(Penn Chinese Treebank, CTB)和北京大学计算语言研究所构造的人

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。

为了充分利用异构数据来提升统计模型的性能, 前人提出几种较为有效的方法^[6-10], 在词法分析任务上均取得不小的性能提升。然而, 这些方法或多或少存在以下 3 个问题: 1) 需要设计复杂的特征模板; 2) 模型参数共享不够完全; 3) 没有同时考虑两个数据规范。

针对前人工作存在的问题, 同时考虑到深度学习方法给词法分析任务带来的积极影响, 本文提出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) 与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) 与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) 与基于传统离散特征的 CRF 耦合模型相比, 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 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。因此, 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(尤其在中文处理任务中)。

Jiang 等^[6]首先提出类似堆叠学习^[11]的指导特征(guide-feature)方法, 用于在 CTB 和 PD 数据集上的分词词性联合标注任务。Sun 等^[12]进一步扩展指导特征方法, 并提出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 等^[9]提出的耦合模型能够在两个数据集上都使用联合特征, 利用联合捆绑标签训练模型参数, 更好地融合两个异构数据集之间的共性。

Qiu 等^[7]提出线性耦合模型, 但他们的模型只使用单独标签特征。Li 等^[9]提出的耦合模型类似于因子 CRF^[13], 从某种意义上说, 捆绑标签可以分解为两个连接的潜在变量, 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, 已有很多关于模糊标注的研究用于分类任务^[14]、序列标注任务^[15]以及分析任务^[16-17]。Li 等^[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。在之前的模型^[9]基础上, Li 等^[10]对数据的可能标签集合进行修剪, 在更小更精确的标签空间上进行建模。

此外, 异构数据的使用也与多任务学习密切相关。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 等^[8]提出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, 单独计算两个数据各自的标签得分, 并预测最终标签序列。

2 基于双向长短期记忆网络的标注模型

依据目前词法分析中基于深度学习的主流方法, 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 4 个部分。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。

2.1 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

表 1 异构数据举例
Table 1 An 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	别	是	我	国	经	济	下	滑	。
	B@AD	I@AD	E@AD	S@PN	S@NN	B@NN	E@NN	B@VV	E@VV	S@PU
PD-分词词性	特	别	是	我	国	经	济	下	滑	。
	B@d	E@d	S@v	B@n	E@n	B@n	E@n	B@v	E@v	S@w

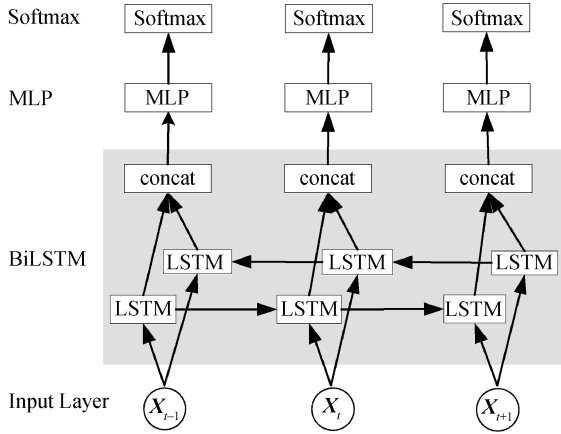


图 1 长短期记忆网络模型
Fig. 1 BiLSTM model

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 $S = \{w_1, w_2, w_3, \dots, w_n\}$, 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 $w_i = \{c_{i,1}, c_{i,2}, c_{i,3}, \dots, c_{i,m}\}$, $c_{i,j}$ 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 h_{lm} 和 h_{rm} 拼接到 w_i 对应的词向量 e_w 后面, 得到 w_i 的最终特征向量 X_i :

$$X_i = e_w \oplus h_{lm} \oplus h_{rm}. \quad (1)$$

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 $S = \{c_1, c_2, c_3, \dots, c_n\}$, 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, 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 X_i 可以表示为

$$X_i = e_{c_{i-1}c_i} \oplus e_{c_i}, \quad (2)$$

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

2.2 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 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 h_i , 形式上可以定义为

$$i_i = \sigma(W_{in} \cdot [h_{i-1}, X_i] + b_{in}), \quad (3)$$

$$f_i = \sigma(W_{fg} \cdot [h_{i-1}, X_i] + b_{fg}), \quad (4)$$

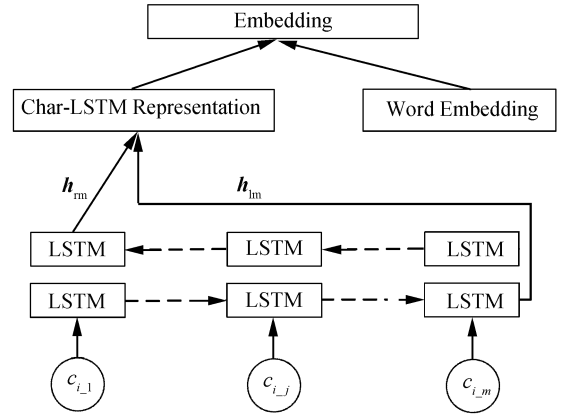


图 2 词的确切表示
Fig. 2 Representation of word

$$o_i = \sigma(W_{out} \cdot [h_{i-1}, X_i] + b_{out}), \quad (5)$$

$$c_i = f_i \cdot c_{i-1} + i_i \cdot \tanh(W_c \cdot [h_{i-1}, X_i] + b_c), \quad (6)$$

$$h_i = o_i \cdot \tanh(c_i), \quad (7)$$

其中, i_i, f_i, o_i, 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, X_i 和 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, W 和 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, 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 h_i :

$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i}, \quad (8)$$

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

2.3 得分层

得分层使用 MLP 来计算每个标签的得分, 将 BiLSTM 的输出 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 Score_i :

$$\text{Score}_i = W_{mlp} \cdot h_i \oplus b_{mlp}, \quad (9)$$

其中, W_{mlp} 和 b_{mlp} 分别表示 MLP 层的权重和偏置。

2.4 预测层

我们尝试对比 CRF 和 local loss, 发现性能差别不大, 并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, 并将交叉熵 (CrossEntropy) 作为目标函数。

对于给定词 w_i , 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

$$y'_j = \text{Softmax}(\text{Score}_j) = \frac{e^{\text{Score}_j}}{\sum_{k=1}^m e^{\text{Score}_k}}, \quad (10)$$

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

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

其中 y_j 表示期望的概率分布。

3 基于多任务学习的异构序列标注数据利用

为了有效地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 Chen 等^[8]提出的基于神经网络多任务学习的异构数据模型。

与基准模型相比, 多任务学习方法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, 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, 在扩大训练语料的基础上提升基准模型的性能。

多任务学习方法与基准模型的网络框架基本上一致, 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

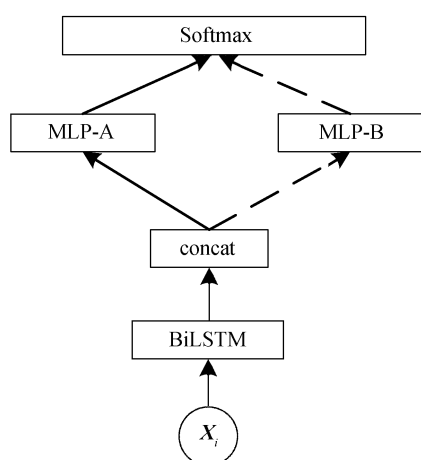


图 3 多任务学习模型

Fig. 3 Multi-Task Learning model

MLP-B 中计算标签得分。模型相对简单, 当输入只有单一语料时, 模型本身相当于基准模型。

4 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集上的标签序列, 只有输入层和 BiLSTM 层同时由两个数据集训练得到, MLP 层单独训练, 所以与 Li 等^[9]提出的耦合模型相比, 异构数据集训练的参数较少。

Li 等^[9]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(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 和图 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 3 个 MLP 中, 分别为计算两个规范上独立标签得分的 MLP-A 和 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

$$\begin{aligned} & \text{Score}(s, i, [t^a, t^b]) \\ &= \text{Score}_{\text{joint}}(s, i, [t^a, t^b]) + \text{Score}_{\text{sep}_a}(s, i, t^a) + \\ & \quad \text{Score}_{\text{sep}_b}(s, i, t^b), \end{aligned} \quad (12)$$

其中, $\text{Score}_{\text{joint}}(s, i, [t^a, t^b])$ 表示句子 s 中第 i 个词被

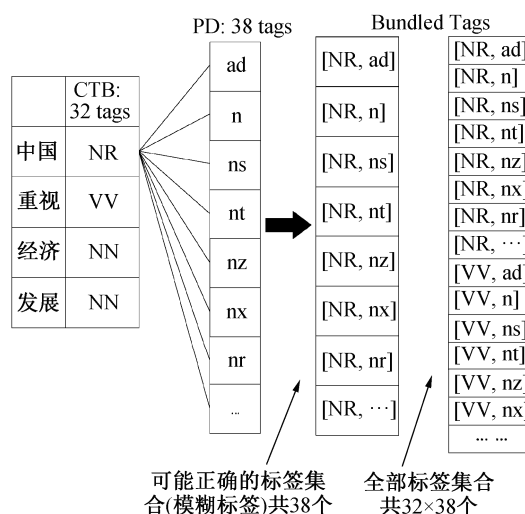


图 4 词性模糊标签示例

Fig. 4 An example of POS ambiguous labeling

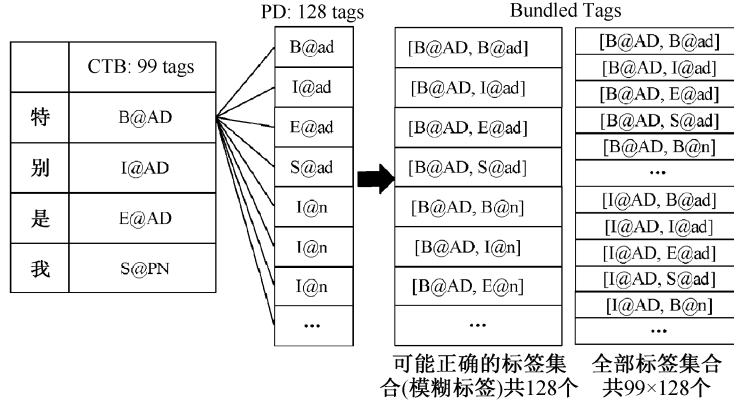


图 5 分词词性联合模糊标签示例

Fig. 5 An example of WS&POS ambiguous labeling

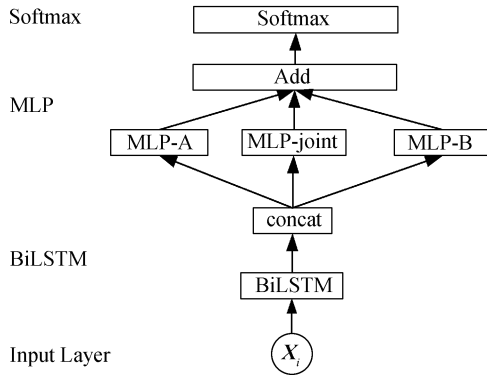


图 6 神经耦合模型

Fig. 6 Neural coupled model

标为联合标签 $[t^a, t^b]$ 的得分, $\text{Score}_{\text{sep}_a}(s, i, t^a) + \text{Score}_{\text{sep}_b}(s, i, t^b)$ 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 以及 B 语料标签集合中独立标签 t^b 的得分。

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扩展。例如 CTB5 的词性标签数量为 $|T_{\text{CTB5}}|=32$, PD 的词性标签数量为 $|T_{\text{PD}}|=38$, 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

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

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。

5 实验

5.1 实验设置

我们在词性标注任务和分词词性联合标注任务上分别进行大量的实验来验证本文方法的有效性。

数据集设置 采用 3 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, 这 3 个数据集分别为 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 等(2016)^[10]和 Li 等(2015)^[9]提供, 简称 PD1 和 PD2。由于缺少 Chen 等^[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 Li 等^[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。在 1000 句人工标注数据中, 共有 27942 个词标有 PD 规范的词性标签, 只有 5769 个词人工标有 CTB 规范的标签。

评价指标 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) 和综合指标值 (F-measure, F)。其中, 分词词性联合的评价不仅仅根据分词的 B, I, E, S 标签计算 P , R 和 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

表 2 数据规模统计
Table 2 Data size statistics

语料		句数	词数
CTB	train	16091	437911
	dev	803	20454
	test	1910	50319
PD1	train	46815	1097839
	dev	2000	46182
	test	5000	118174
PD2	train	273883	6499208
	dev	1000	23427
	test	2500	58301
newly labeled		1000	标注 CTB 词性 标注 PD 词性
			5769 27942

用 Li 等^[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,并对其中人工标注 CTB 规范词性的 5769 个词评价词性准确率。

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^[18]对模型进行更新。为了合理地混合两个异构数据,我们采用 Li 等^[19]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 和 N 个句子混合训练,在各自的 dev 数据上词性准确率或分词词性标注的 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 和 PD1 的训练集进行实验对比后,采用 $M=N=16091$ (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。

5.2 词性标注结果

为了验证神经耦合模型的有效性,首先在词性标注任务上做实验,与前人工作进行对比。为了更好地与 Li 等(2015)^[9]和 Li 等(2016)^[10]的结果进行比较,我们在 CTB5, PD2 和 CTB5, PD1 两组异构数据集上分别训练 3 个模型,结果如表 3 和表 4 所示。

在基准词性标注准确率上, BiLSTM 模型要比

表 3 不同方法在 CTB5 和 PD1 上的词性标注准确率(%)
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 等(2016) ^[10]	Baseline CRF	—	94.07	—
	Coupled CRF	—	94.83	—

说明:粗体数字表示最佳结果,下同。

表 4 不同方法在 CTB5 和 PD2 上的词性标注准确率(%)
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 等(2015) ^[9]	Baseline CRF	94.28	94.10	—
	Coupled CRF	95.10	95.00	—

基于传统离散特征的 CRF 模型略高。在利用异构数据之后,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,准确率都比基准模型有明显的提高。尤其是对数据规模相对较小的 CTB5 数据集,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%,神经耦合模型提升 1.12% 和 1.63%。结果表明使用异构数据确实能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 数据集上,神经耦合模型的结果基本上与之持平,甚至在 PD1 上还略低。在 CTB5 的测试集上,尤其是与 PD1 组成异构数据集进行的实验中,神经耦合模型的准确率比多任务学习方法高出 0.40%。这是由于 CTB5 数据集规模较小,PD 数据集规模较大,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。PD 能从 CTB5 中学到的很少,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,在两组异构数据集实验中,基于深度学习的耦合模型的准确率有很明显的提高。除在 CTB5 上分别有 0.62% 和 0.96% 的提升外,即使是在 PD1 上,也有 0.29% 的提升。说明在词性标注任务上,深度学习方法确实比传统离散特征方法更具备优势。也进一步验证了本文提出的深度学习耦合模型能更充分地利用异构数据。

在 PD1 和 PD2 上,几个模型的性能差别不明显,无论是多任务学习方法还是耦合标注方法,都不能有效地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。

5.3 分词词性联合标注结果

为进一步验证模型的有效性,我们在分词词性联合标注任务上做了一组对比实验。为与 Li 等^[10]的实验结果对比,我们采用同样的数据集 CTB5 和 PD1,并训练基准模型、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。

与词性标注结果不同的是,与传统离散特征耦合 CRF 模型相比,神经耦合模型的性能基本上与之持平。与基于传统离散特征的基准 CRF 相比,基准 BiLSTM 模型的性能相当,在 CTB5 测试集上略低,在 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 基本上表现一致, 且 BiLSTM 较容易受数据规模的影响。

与多任务学习方法相比, 神经耦合模型的性能也没有明显的提升, 基本上持平。在分词任务上, 相对于多任务学习方法, 耦合模型并无明显优势。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, 准确率比基准模型均有 1.2% 左右的明显提升, 表明异构数据模型确实能够有效地提升分词性能。

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 规模比 CTB 大, 本身性能已较高, 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 等^[10]的实验结果均表明这一点。由于实验受到随机初始化以及 batch 大小等因素的影响, 各模型的性能会有轻微的浮动。

5.4 模型的异构数据词性转化

我们对模型对异构数据的词性转化进行实验分析, 结果如表 6 所示。

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 训练得到, 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明显优于传统离散特征方法, 取得 0.90% 的性能提升。

在利用 CTB5 和 PD1 异构数据训练的模型上, 与基准模型相比, 神经耦合标注模型和多任务学习

表 5 不同方法在 CTB5 和 PD1 上的分词词性联合标注结果 (%)

Table 5 The result 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 等 (2016) ^[10]	Baseline CRF	89.60	89.38	89.49	92.74	92.20
	Coupled CRF	90.68	90.49	90.58	92.70	92.19

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

Baseline CRF	90.59
Coupled CRF (with PD2)	93.90

方法都有很大程度提升, 分别提升 1.61% 和 1.87%。在利用 CTB5 和 PD2 异构数据训练的模型上, 不仅多任务学习方法的转化性能进一步提升, 而且基于深度学习和基于传统离散特征的异构数据模型相对于基准模型的提升也都更加明显, 表明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, 无论在 CTB5 和 PD1 数据集上, 还是在更大规模的 CTB5 和 PD2 数据集上的实验结果都表明, 由于神经耦合模型更全面地考虑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。并且, 在较大规模的异构数据集上, 神经耦合模型的准确率比多任务学习方法有 0.38% 的提升。

与 Li 等^[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性能上有 0.64% 的提升, 达到各模型的最佳。结果表明, 基于深度学习方法的耦合模型能够更好地完成异构数据的词性转化与融合。

5.5 实验结果分析

相同模型在不同数据集上的差异表明, 在异构的两组数据上, 规模更大的 CTB5 和 PD2 能够取得更多的性能提升。在异构数据内部, 规模较大的 PD 数据集并不能从 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, 也说明耦合模型能够更充分利用异构数据来提升统计模型性能词性标注的准确率。

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地考虑异构规范信息, 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,从而进行更好的数据融合。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等^[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。

6 结语

本文借鉴前人的工作,将耦合模型扩展到深度学习框架中。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。与前人工作的对比表明,本文模型在词性标注任务上的性能更好,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捆绑标签得分,更充分地融入不同的规范信息,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 // Proceedings of 43rd ACL. Ann Arbor, 2005: 91–98
- [2] McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation // Proceedings of 17th ICML. Stanford, 2000: 591–598
- [3] Berger A L, Pietra S A D, Pietra V J D. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996, 22(1): 39–68
- [4] Xue Naiwen, Xia Fei, Chiou Fudong, et al. The Penn Chinese TreeBank: phrase structure annotation of a large corpus. *Natural Language Engineering*, 2005, 11(2): 207–238
- [5] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范. *中文信息学报*, 2002, 16(5): 49–64
- [6] Jiang Wenbin, Huang Liang, Liu Qun. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging: a case study // Proceedings of 47th ACL. Singapore, 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 // Proceedings of EMNLP. Seattle, 2013: 658–668
- [8] Chen Hongshen, Zhang Yue, Liu Qun, et al. Neural network for heterogeneous annotations // Proceedings of EMNLP. Austin, 2016: 731–741
- [9] Li Zhenghua, Chao Jiayuan, Zhang Min, et al. Coupled sequence labeling on heterogeneous annotations: POS tagging as a case study // Proceedings of 53rd ACL. Beijing, 2015: 1783–1792
- [10] Li Zhenghua, Chao Jiayuan, Zhang Min, et al. Fast coupled sequence labeling on heterogeneous annotations via context-aware pruning // Proceedings of EMNLP. Austin, 2016: 753–762
- [11] Nivre J, McDonald R. Integrating graph-based and transition-based dependency parsers // Proceedings of 46th ACL. Columbus, 2008: 950–958
- [12] Sun Weiwei, Wan Xiaojun. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations // Proceedings of 50th ACL. Jeju Island, 2012: 232–241
- [13] 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
- [14] Jin Rong, Ghahramani Z. Learning with multiple labels // Proceedings of NIPS. Vancouver, 2002: 921–928
- [15] Dredze M, Talukdar P P, Crammer K. Sequence learning from data with multiple labels // Proceedings of the ECML/PKDD-09 Workshop on Preference Learning. Bled, 2009: 39–48
- [16] Riezler S, King T H, Kaplan R M. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques // Proceedings of 40th ACL. Philadelphia, 2002: 271–278
- [17] Täckström O, McDonald R, Nivre J. Target language adaptation of discriminative transfer parsers // Proceedings of NAACL. Atlanta, 2013: 1061–1071
- [18] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2017–01–30) [2018–12–08]. <https://arxiv.org/abs/1412.6980>
- [19] Li Zhenghua, Zhang Min, Chen Wenliang, et al. Ambiguity-aware ensemble training for semi-supervised dependency parsing // Proceedings of 52nd ACL. Baltimore, 2014: 457–467