北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.037

面向词性标注的多资源转化研究

高恩婷[1]　巢佳媛[2]　李正华[2,†]

1.苏州科学技术学院电子与信息工程学院，苏州 215011; 2.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　提出利用多资源转化方法进行词性标注研究，旨在将源端资源的标注进行转化，以符合目标端标注规范，进而将转化后的资源和目标资源合并，增大训练数据规模。作者做了两方面创新：在转化过程中，额外利用指导特征的置信度信息；在转化后的资源中，用模糊标注表示方法减少错误标注。实验表明，利用置信度信息能有效帮助转化，而模糊标注表示方法的影响不大。

关键词　词性标注转化；条件随机场；词性标注
中图分类号　TP391

Conversion of Multiple Resources for POS Tagging

GAO Enting[1]　CHAO Jiayuan[2]　LI Zhenghua[2,†]

1. College of Electronics & Information Engineering, Suzhou University of Science and Technology, Suzhou 215011;
2. School of Computer Science & Technology, Soochow University, Suzhou 215006;
† Corresponding author, E-mail: zhli13@suda.edu.cn

Abstract　The authors propose an annotation conversion method using multiple resources for POS tagging, aiming to convert the source-side annotations into target-side and then combine the data to get larger training data. Two innovate strategies are proposed. The first strategy uses reliability information of guide features. The second strategy uses ambiguous labelings to improve the quality of converted data. Results demonstrate that the first strategy is helpful for annotation conversion while the second does little to conversion.

Key words　annotation conversion; conditional random field; POS tagging

目前有指导的统计机器学习方法通常使用单个标注资源训练模型参数，但是单个标注资源规模小、领域覆盖面窄。目前常用的分词、词性标注和句法分析语料为宾州中文树库 Penn Chinese Treebank5.1[1] (CTB5)，仅约 2 万句，语料规模小，而人工标注新的语料来扩大语料规模的方式费事费力。另一方面，不同标注规范的其它资源的存在，为研究界提出了一个新的问题和机遇，即如何在统计模型中利用多种不同标注规范的资源，以提高分析效果。例如，北京大学标注的分词词性标注语料(即人民日报语料 People's Daily[2](PD))，约 30 万句；对句法分析而言，还有清华大学的短语树库(TCT)和哈工大的依存树库(CDT)。

近年来，研究者们尝试额外利用遵守不同标注规范的源端资源，提高模型在目标资源上的分析效果。其中，最具代表性的方法为 Stacked Learning，即基于指导特征的方法，最早由 Jiang 等[3]应用于分词和词性标注联合任务。其基本思路是：1) 在源端资源上训练一个源端标注器(level-0 tagger)；2) 利用源端标注器，自动分析所有目标端资源(包括训练和测试数据)，这样，目标端资源上就额外增加了自动产生的包含噪声的源端标注结果；3) 训练目标端标注器(level-1 tagger)，将源端标记作为额外指导特征加入到模型中，利用这种指导特征帮助模型预测；4) 利用训练好的目标端标注器，分析测试数据，同样将源端标注结果作为指导特征。

收稿日期: 2014–06–30; 修回日期: 2014–10–27; 网络出版时间: 2014-12-01 09:26

基于指导特征的方法在分词、词性和依存句法分析任务上取得了很好的效果, 然而存在两方面的问题。1) 源端资源不直接参与目标端模型的训练过程, 而是以指导特征的形式间接指导目标模型。目标端模型在训练阶段只使用目标端资源中的句子, 受限于目标端资源中出现的语言现象。2) 实际应用中需要两次解码, 效率较低。每个句子需要接受两次标注, 第一次解码时由源端标注器产生指导信息, 第二次解码则使用带指导特征的目标端词性标注器。

为了解决基于指导特征的方法存在的两个问题, 本文以词性标注为案例, 将基于指导特征的方法应用于标注规范转化任务, 旨在将源端资源的标注进行转化, 以符合目标标注规范, 进而将转化后的资源和目标资源合并, 增大训练数据规模。我们将这种方法称为基于转化的方法。具体方法上, 将源端资源作为测试数据, 利用基于指导特征的模型进行自动分析, 从而得到符合目标标注规范的含有噪音的结果。本文做了两方面创新尝试。1) 在基于指导特征的方法中, 额外利用指导特征的置信度信息。即 level-1 tagger 将 level-0 tagger 产生的概率信息作为指导特征的置信度, 增加到模型训练中。实验表明利用置信度信息可以帮助基于指导特征的方法, 帮助基于转化的方法。2) 转化后资源包含的错误会影响模型训练过程。为了尽量削弱这种影响, 本文利用两种不同的转化方法(一种使用指导特征的置信度, 另一种不使用置信度)对源端数据进行转化, 将得到的结果合并, 形成模糊标注。换言之, 如果两种转化方法在某一个词上产生的标记不一致, 那么两个标记都保留下来, 即允许训练数据中的某些词同时拥有多个不同的词性标记。然而实验结果表明, 这种方法并不能帮助基于转化的方法。

本文将 PD 作为源端资源, CTB5 作为目标端资源, 进行深入实验和分析。结果表明, 本文提出的方法能有效提高词性标注性能。与传统的基于单个资源的方法相比, 基于转化的方法可以将词性标注准确率从 94.11%提高到 95.07%。

1 相关工作

近年来, 单纯的汉语词性标注方面的研究工作比较少, 进展比较缓慢。Tseng 等[4]提出用丰富的词性分析来帮助未登录词的词性标注。Huang 等[5]使用自学习(Self-training)来提高词性标注性能。还有学者尝试利用句法信息来帮助词性标注: Huang 等[6]使用短语句法分析结果来帮助词性标注, Li 等[7]则利用依存句法分析结果来帮助词性标注。进而, 有研究提出词性标注和句法分析联合模型[8–10], 允许词性标记和句法信息互相交互和帮助。

在利用多个资源方面, Jiang 等[3]首先提出基于指导的方法, 提高分词和词性标注的准确率。Sun 等[11]提出一种利用结构化指导特征的方法, 利用不同标注规范的分词资源。Zhu 等[12]和 Li 等[13]将类似的想法应用到句法分析上。Qiu 等[14]提出一种联合学习方法, 在两种资源的基础上同时学习两种不同标注规范的分词词性标记。

与本文最接近的两种方法分别为 Meng 等[15]和 Jiang 等[16]。Meng 等直接将基于指导特征的方法应用于资源转化, 以提高分词和词性标注联合任务的准确率。Jiang 等面向分词任务, 从两个方向多次使用基于指导特征的方法进行资源转化, 在迭代中不断提高转化后资源的质量。与这两个工作不同, 本文对基于指导特征的方法进行了改进, 额外使用指导特征的置信度, 并且通过实验说明其有效性。另外, 本文还初步尝试了基于模糊标注的方法以控制转化后资源中包含的噪声。

“模糊标注”的思想又称为 ambiguous labelings 或 multiple labelings, 在分类[17]、序列标注[18–19]、和依存句法分析[20]上都有应用。目前, 本文在模糊标注使用方式上还比较简单, 未来会尝试提出更加有效合理的方法。

2 基于 CRF 的词性标注模型

词性标注是一个典型的序列标注问题, 常用的方法有最大熵(Maximum Entropy)模型[21]、基于条件随机场(Conditional Random Fields, CRF)模型[22]和感知器(Perceptron)模型[23]。CRF 结合最大熵模型和隐马尔科夫模型的特点, 也是近年来处理序列标注问题效果好的模型。同时, 由于感知器模型为线性模型, 无法获得概率信息, 而本文多处会用到概率信息, 因此我们利用基于 CRF 模型进行词性标记转化。

2.1 模型定义

给定输入句子 $\boldsymbol{x}=w_1, \ldots, w_i, \ldots, w_n$, 对应的词性标注序列为 $\boldsymbol{t} = t_1 \ldots t_i \ldots t_n$ ($t_i \in T$), T 为标注规范中包含的词性标记集合。词性标注旨在给定输入句子

$\boldsymbol{x}$, 输出概率最优的词性标注序列 $\boldsymbol{t}^*$, 如式(1):

$$\boldsymbol{t}^* = \text{argmax}_{\boldsymbol{t}\in Y(x)} P(\boldsymbol{x}, \boldsymbol{t} \mid \boldsymbol{\theta}) \quad (1)$$

其中, $Y(\boldsymbol{x})$ 表示输入的句子 $\boldsymbol{x}$ 对应的所有词性标记序列路径的集合, 即搜索空间; θ表示模型参数, 即特征权重向量。

CRF 属于对数线性(Log-linear)模型, 它将一个词性序列的概率定义如下:

$$\text{p}(\boldsymbol{t} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp\{\text{Score}_{\text{pos}}(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{\theta})\}}{Z(\boldsymbol{x}; \boldsymbol{\theta})}, \quad (2)$$

$$Z(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{t\in Y(x)} \exp\{\text{Score}_{\text{pos}}(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{\theta})\}, \quad (3)$$

其中, $\text{Score}_{\text{pos}}(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{\theta})$ 为一个词性序列的分值, $Z(x; \theta)$ 为正则化因子。我们采用一元(unigram)特征 $f_{\text{pu}}(\boldsymbol{x}, \boldsymbol{t}_i)$和二元(bigram)特征 $f_{\text{pb}}(\boldsymbol{x}, \boldsymbol{t}_{i\text{-}1}, t_i)$。因此词性序列的分值可以定义如下:

$$\text{Score}_{\text{pos}}(x, t; \theta) = \theta \cdot (x, t) = \sum_{i=1}^{n} \{\boldsymbol{\theta}_{\text{pu}} \cdot \boldsymbol{f}_{\text{pu}}(x, t_i) + \boldsymbol{\theta}_{\text{pb}} \cdot \boldsymbol{f}_{\text{pb}}(x, t_{i-1}, t_i)\}。 \quad (4)$$

我们采用 Li 等[13]词性标注特征集合, 如表 1 所示。其中, o 为字符串连接符, $c_{i,k}$为w_i的第 k 个字符, $c_{i,0}$是第一个字符, $c_{i,-1}$是最后一个字符, c_i为 w_i 中包含的字符数, prefix/suffix(w_i, k)表示 w_i 的长度为 k 的前缀/后缀。

表 1 词性标注特征集合

Table 1 Features for POS tagging

POS Unigram Features: $f_{\text{pu}}(x,i,t_i)$
01 $t_i \circ w_i$
02 $t_i \circ w_{i-1}$
03 $t_i \circ w_{i+1}$
04 $t_i \circ w_i \circ c_{i-1,-1}$
05 $t_i \circ w_i \circ c_{i+1,0}$
06 $t_i \circ c_{i,0}$
07 $t_i \circ c_{i,-1}$
08 $t_i \circ c_{i,\text{k}}, 0 < k < \#c_i - 1$
09 $t_i \circ c_{i,0} \circ c_{i,\text{k}}, 0 < k < \#c_i - 1$
10 $t_i \circ c_{i,-1} \circ c_{i,\text{k}}, 0 < k < \#c_i - 1$
11 if $\#c_i = 1$ then $t_i \circ w_i \circ c_{i-1,-1} \circ c_{i+1,0}$
12 if $c_{i,k} = c_{i,k+1}$ then $t_i \circ c_{i,k} \circ$ "consecutive"
13 $t_i \circ \text{prefix}(w_i, k), 1 \leqslant k \leqslant 4, k \leqslant \#c_i$
14 $t_i \circ \text{suffix}(w_i, k), 1 \leqslant k \leqslant 4, k \leqslant \#c_i$
POS Bigram Features: $f_{\text{pb}}(x,i,t_{i-1},t_i)$
15 $t_i \circ t_{i-1}$

2.2 模型训练

假设人工标注的训练数据为 $D = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{N}$。则训练数据 D 的对数似然函数以及对其求偏导后的梯度如式(5)和(6):

$$\text{L}(\boldsymbol{D}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(t_i \mid x_i; \boldsymbol{\theta}) \quad (5)$$

$$\frac{\partial L(\boldsymbol{D}; \boldsymbol{\theta})}{\partial \theta} = \sum_{i=1}^{N} \{f(x_i, t_i) - \sum_{t\in Y(x_i)} P(\boldsymbol{t} \mid x_i; \theta) f(x_i, \boldsymbol{t})\} \quad (6)$$

式(6)等号左边的第一项为正确词性序列获得的特征的统计数(Empirical Count), 第二项为模型期望(model expectation)。由于 $Y(\boldsymbol{x}_\text{i})$中包含了指数级的词性序列, 直接遍历这个集合是不实际的, 因此, 我们采用经典的 Forward-Backward 动态规划算法, 在多项式时间内, 计算特征的模型期望。利用随机梯度下降(Stochastic Gradient Descent, SGD)训练算法得到模型参数, 即特征权重向量。

3 基于指导特征的方法

3.1 基本框架

图 1 为基于指导特征的方法的框架图。给定源端数据(Source Data) S=$\{(x_i, t_i)\}i$, 目标端数据(Target Data) T=$\{(x_j, t_j)\}j$, 在源端数据上训练一个词性标注器(Tagger-S), 用它对目标端数据进行词性标注, 产生源端自动词性标注。得到标注后的数据(Tagged data): $T^S = \{(x_j, t_j^S)\}_j$, 即第 j 个句子 x_j 的词性标注序列为 t_j^S。为将源端自动词性作为额外的指导特征, 添加到目标端词性标注器中, 我们进行了数据处理, 得到源端词性标注的目标端数据(Target data with source annotation) T^{+S} = $\{(x_j, t_j, t_j^S)\}_j$, 即目标端数据中第 j 个句子 x_j同时有两个词性标注序列 t_j 和 t_j^S。在该数据上训练得到的词性标注器就是带有指导特征的目标端词性标注器。

除了表 1 中使用的基本词性特征集合之外, 带有指导特征的目标端词性标注器还用到额外的基于源端词性标记的指导特征, 见表 2。

表 2 额外的指导特征

Table 2 Additional guide features

Guide POS Features: $\text{f}_g(\text{x}, \text{t}^S, i, t_i)$	
01 $f_{\text{pu}}(x,i,t_i) \circ t_i^S$	05 $t_i \circ t_{i-1}^S \circ t_i^S$
02 $t_i \circ t_i^S$	06 $t_i \circ t_i^S \circ t_{i+1}^S$
03 $t_i \circ t_{i+1}^S$	07 $t_i \circ t_{i-1}^S \circ t_{i+1}^S$
04 $t_i \circ t_{i-1}^S$	08 $t_i \circ t_{i-1}^S \circ t_i^S \circ t_{i+1}^S$

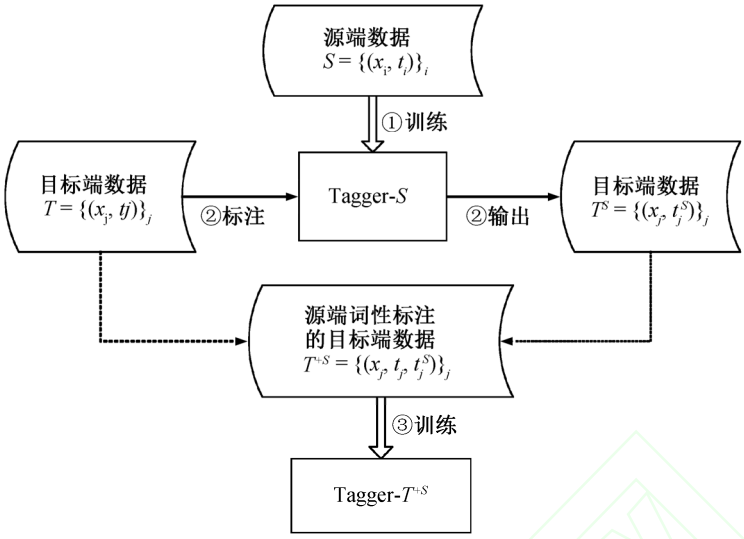

图 1　基于指导特征的方法框架图

Fig. 1　Framework for our guide feature based method

如表 3 中给出的输入词序列“六月 在 浦东 正式 成立”, 利用基于指导特征的方法, 对其词性标注过程分为两步: 1) 用源端词性标注器 Tagger-S 分析输入词序列, 得到源端词性标记, 即指导特征(表中第 2 行); 2) 利用带指导特征的目标端词性标注器 Tagger-T^{+S} 分析带指导特征的该词序, 得到目标端词性 A(表中第 4 行)。

3.2　指导特征的置信度

一个特征的可靠性可以通过其置信度表示出来。以前的研究在使用指导特征时, 没有考虑到指导特征本身的置信度。为此我们将指导词性的边缘概率作为指导特征的置信度, 加入到目标端词性标注器。在目标端数据分析时, 产生源端自动词性标注的同时也产生对应的边缘概率, 即指导特征的置信度, 然后将得到的源端自动词性标注作为额外的特征, 以及其指导特征置信度, 添加到目标端数据中, 最后得到“带指导特征置信度的目标词性标注器”。

表 3　基于指导特征的方法和基于转化的方法示例

Table 3　An example of guide features based method and conversion method

输入源端词序列	六月	在	浦东	正式	成立
指导特征	t	p	nr	ad	v
指导特征置信度	0.8	0.9	0.6	0.8	0.9
目标端词性(A)	NT	P	**NN**	AD	VV
目标端词性(B)	**NN**	P	NR	AD	VV
模糊标注表示方法	NT_NN	P	NN_NR	AD	VV

带指导特征置信度的目标词性标注器使用指导特征时, 将源端词性标记的边缘概率作为特征值(一般特征对应的特征值都是 0/1 值)。例如, 表 2 中的 01 特征 $t_i^S \circ t_i$, 将 t_i^S 的边缘概率作为该特征的特征值。当某个指导特征只用到多个源端词性标记时(如 08 特征), 使用多个边缘概率的最小值。

给定输入词序列“六月 在 浦东 正式 成立”, 首先用源端词性标注器 Tagger-S 分析输入词序, 得到源端词性标记同时产生词性对应的边缘概率, 即指导特征(表 3 第 2 行)和指导特征置信度(表 3 第 3 行); 然后利用带指导特征的目标端词性标注器 Tagger-T^{+S}, 分析带指导特征和指导特征置信度的词序列, 得到目标端词性 B(表 3 第 5 行)。

4　基于转化的方法

4.1　基本框架

图 2 给出基于转化的方法框架图。给定源端数据, 经过带指导特征的目标端词性标注器分析后, 产生目标端自动词性标注, 得到带目标端词性标注的源端数据 $S^T=\{(x_i, t_i^T)\}_i$, 即第 i 个句子的词性标注序列为 t_i^T。此时, 我们就实现了将源端数据词性标记转化为目标端词性标记的过程, 随后, 可以用源端数据和目标端数据合并为更大规模的训练集, 训练一个新的词性标注器 Tagger-(T∪S)。本文采用基于指导特征的方法进行词性标记转化, 即分别利用“带指导特征的目标端词性标注器”和“带指导

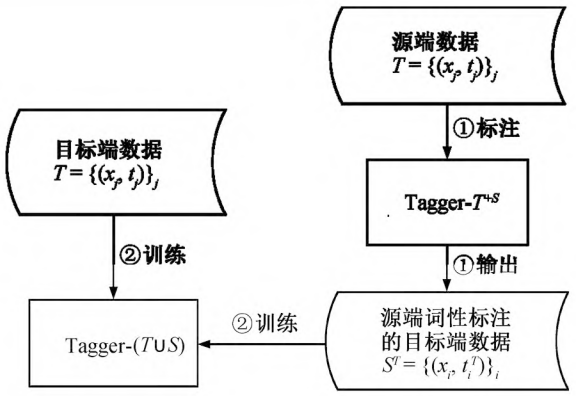

图 2　基于转化的方法框架图

Fig. 2　Framework for our conversion method

特征置信度的目标端词性标注器”实现转化。以表 3 中的输入源端词序列为例, 带指导特征的目标端词性标注器分析带指导特征的该词序列, 得到目标端词性 A(表 3 第 4 行), 带指导特征置信度的目标端词性标注器分析带指导特征和指导特征置信度的该词序列, 得到目标端词性 B(表 3 第 5 行)。

4.2　模糊标注的训练过程

转化后的语料难免存在错误的词性标记, 为降低错误标记对词性标注系统的影响, 本文利用“模糊标注”的表示方法, 对两种转化后的词性标记进行模糊处理, 即允许训练数据中的某些词同时拥有多个不同的词性标记。

以表 3 给出的源端词序列为例, 转化后的目标端词性 A 和 B 中存在错误的词性标注, 目标端词性 A 中“浦东”标注的词性有误而目标端词性 B 中“六月”的词性有误(在表中以黑色表示)。因此, 我们采用模糊标注的表示方法将转化后的词性标记进行模糊处理。如果标注的目标端词性 A 和 B 不一致, 那么两个标记都保留下来, 如果两词性一致, 则维持标注词性不变(表 3 第 6 行)。然后, 将模糊标注方法处理后的源端数据和目标端数据合并为更大规模的训练集, 训练一个新的词性标注器。

为了允许模型利用模糊标注表示方法训练语料, 我们对基于 CRF 的词性标注模型进行扩展。假设包含模糊标记的训练实例的集合为 $\boldsymbol{D}'=\{(\boldsymbol{u}_i, V_i)\}_{i=1}^{M}$, 其中$V_i$表示模糊标注中包含的所有可能的词性序列集合, 则 $\boldsymbol{D}'$的对数似然函数为

$$L(\boldsymbol{D}';\boldsymbol{\theta})=\sum_{i=1}^{M}\log\{\log\{\sum_{t\in V_i}p(\boldsymbol{t}\mid\boldsymbol{u}_i;\boldsymbol{\theta})\}\} \qquad (7)$$

即一个模糊标注的训练实例的概率为所有可能的词性序列的概率之和, 同样, 可以求解似然关于特征权重向量$\boldsymbol{\theta}$ 的偏导[18]。

5　实验结果和实验分析

5.1　实验数据

本文在北京大学 1998 年上半年《人民日报》PD 语料和宾州中文树库 CTB5 语料上做了详细的实验和分析, 其中, PD 为源端数据, CTB5 为目标端数据。本文采用词性标注准确率(即测试集上正确的词性标注所占的比例), 来评价系统的性能。表 4 为语料的统计数据。

5.2　实验结果

首先, 我们在 CTB5 语料上训练了一个词性标注器, 以此为基准系统。然后, 根据本文提出的方法, 做了多组实验, 在 CTB5 开发集和测试集上的实验结果如表 5 所示。表中, PD_A 是利用基于指导特征的方法(无置信度 A)建立的词性标注器转化的语料; PD_B 是利用基于指导特征的方法(有置信度 B)转化后的语料; 而 $PD_{A\cup B}$ 是利用模糊标注的表示方法合并 PD_A 和 PD_B 后的语料。

实验结果表明, 本文提出的方法能有效提高词性标注准确率。与传统的基于单个资源的方法相比, 基于转化的方法在扩大规模的语料上利用多资源信息, 可以将词性标注准确率从 94.11% 提高到 95.07%; 与此数据集上目前最高的词性标注系统[24]

表 4　语料统计表

Table 4　Corpus statistics

分类	CTB5			PD		
	训练集	开发集	测试集	训练集	开发集	测试集
句数	16091	803	1910	291722	5000	10000
词数	437991	20454	50319	6843978	116887	236528

表 5　实验结果

Table 5　Experiment results

实验方法	训练集	Dev/%	Test/%
基准系统	CTB5	94.24	94.11
基于指导特征的方法 A (无置信度)	CTB5	95.07	94.87
基于指导特征的方法 B (有置信度)	CTB5	95.25	**95.00**
基于转化的方法	CTB5+PD_A	95.23	94.94
	CTB5+PD_B	95.30	95.06
	CTB5+$PD_{A\cup B}$	95.30	**95.07**
词性标注和句法分析联合模型[24]	CTB5	-	94.60

相比，从 94.60%提高到 95.07%。具体而言，本文提出的利用指导特征置信度的方法能够帮助基于指导特征的方法，以及帮助得到较好的转化语料；然而利用模糊标注的表示方法，对基于转化的方法影响不大，原因主要是构成模糊标注的两个模型的差异太小，导致模糊标注中仍然包含不少错误。

5.3 实验分析

为探究本文提出的方法在不同数据规模上对词性标注性能的影响，分别利用基于指导特征的方法(有置信度)和基于模糊标注的转化方法，做了如下实验分析。

1) CTB5 语料规模的影响: PD 语料不变, CTB5 的训练语料规模变为原来的 1/2, 1/4 和 1/8。采用两种方法进行对比实验，结果如图 3 所示。

2) PD 语料规模的影响: CTB5 语料不变, PD 的训练语料规模变为原来的 1/2, 1/4 和 1/8。采用两种方法进行对比实验，结果如图 4 所示。

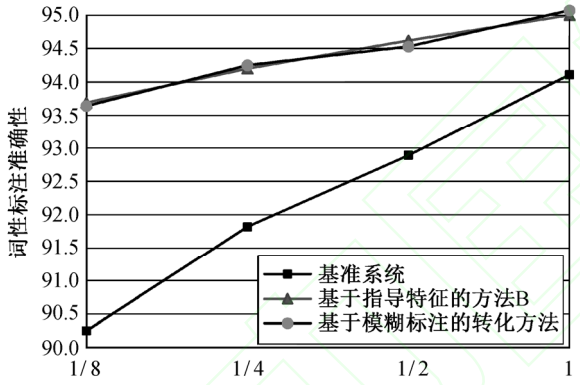

图 3 CTB5 训练集规模的影响

Fig. 3 Effect of training set size of CTB5

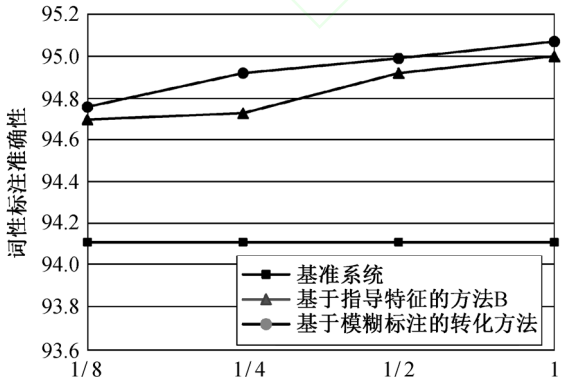

图 4 PD 训练集规模的影响

Fig. 4 Effect of training set size of PD

由图 3 可知，在 PD 语料规模不变的情况下，与基准系统相比，基于指导特征的方法(有置信度)和基于模糊标注的转化方法随着 CTB5 规模变小，对词性标注性能影响越大；而 CTB5 语料规模变化过程中，两种方法对词性标注性能的影响基本一致，但高于基准系统。由图 4 可知，CTB5 语料规模不变的情况下，随着 PD 规模的增大，基于转化的方法的准确率略高于基于指导特征(有置信度)的方法，而与基准系统相比，两方法均能大幅度提高词性标注准确率。

6 总结和展望

针对现有单一资源的词性标注准确率不高的问题，本文利用多资源转化的方法，扩大语料规模领域的覆盖面。本文首次尝试在基于指导特征的方法中考虑指导特征的置信度，并采用模糊标注的方法表示转化后的资源。实验结果显示，基于指导特征的方法可以有效提高词性标注性能，然而模糊标注的方法并不能帮助提高基于转化的方法。我们的未来工作主要包括两个方面: 1) 提出更合理的方法，以利用模糊标注的方法帮助基于转化的方法; 2) 将本文的方法与词性标注和句法分析联合模型相结合，进一步提高资源转化的质量。

参考文献

[1] Xue N, Xia F, Chiou Fudong, et al. The penn chinese treebank: phrase structure annotation of a large corpus. Natural language engineering, 2005, 11(2): 207–238

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

[3] Jiang Wenbin, Huang Liang, Liu Qun. Automatic adaptation of annotation standards: Chinese word segmentation and pos tagging: a case study // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore: Association for Computational Linguistics, 2009: 522–530

[4] Tseng H, Jurafsky D, Manning C. Morphological features help pos tagging of unknown words across language varieties // Proceedings of the fourth SIGHAN workshop on Chinese language processing.

Philadelphia, 2005: 32–39

[5] Huang Zhongqiang, Eidelman V, Harper M. Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-training // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Singapore: Association for Computational Linguistics, 2009: 213–216

[6] Huang Zhongqiang, Harper M P, Wang Wen. Mandarin part-of-speech tagging and discriminative reranking // EMNLP-CoNLL. Prague, 2007: 1093–1102

[7] Li Zhenghua, Che Wanxiang, Liu Ting. Improving chinese pos tagging with dependency parsing // IJCNLP. Chiang Mai, 2011: 1447–1451

[8] Li Zhenghua, Zhang Min, Che Wanxiang, et al. Joint models for Chinese pos tagging and dependency parsing // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 1180–1191

[9] Hatori J, Matsuzaki T, Miyao Y, et al. Incremental joint pos tagging and dependency parsing in Chinese // IJCNLP. Chiang Mai, 2011: 1216–1224

[10] Wang Zhiguo, Xue Nianwen. Joint pos tagging and transition-based constituent parsing in chinese with nonlocal features // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 733–742

[11] Sun Weiwei, Wan Xiaojun. Reducing approximation and estimation errors for chinese lexical processing with heterogeneous annotations // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju: Association for Computational Linguistics, 2012: 232–241

[12] Zhu Muhua, Zhu Jingbo, Hu Minghan. Better automatic treebank conversion using a feature-based approach // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011: 715–719

[13] Li Zhenghua, Liu Ting, Che Wanxiang. Exploiting multiple treebanks for parsing with quasi-synchronous grammars // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju: Association for Computational Linguistics, 2012: 675–684

[14] Qiu Xipeng, Zhao Jiayi, Huang Xuanjing. Joint chinese word segmentation and pos tagging on heterogeneous annotated corpora with multiple task learning // EMNLP. Seattle, 2013: 658–668

[15] Meng Fandong, Xu Jin'An, Jiang Wenbin, et al. A method of merging corpora in different annotation standards: an application on statistics chinese lexical analysis. Journal of Chinese Information Processing, 201: 3–122

[16] Jiang Wenbin, Meng Fandong, Liu Qun, et al. Iterative annotation transformation with predict-self reestimation for chinese word segmentation // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju: Association for Computational Linguistics, 2012: 412–420

[17] Jin Rong, Ghahramani Zoubin. Learning with multiple labels // Advances in neural information rocessing systems, Montreal, 2002: 897–904

[18] Tsuboi Y, Kashima H, Oda H, et al. Training conditional random fields using incomplete annotations // Proceedings of the 22nd International Conference on Computational Linguistics. Manchester: Association for Computational Linguistics, 2008: 897–904

[19] Dredze M, Partha Pratim Talukdar, Crammer K. Sequence learning from data with multiple labels // Workshop Co-Chairs. Slovenia, 2009: 39

[20] Täckström O, McDonald R, Nivre J. Target language adaptation of discriminative transfer parsers // Proceedings of NAACL-HLT 2013. Atlanta, 2013: 1061–1071

[21] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging // Proceedings of the conference on empirical methods in natural language processing. Philadelphia, 1996: 133–142

[22] Lafferty J, McCallum A, Fernando CN Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data // Proceedings

of the 18th International Conference on Machine Learning. Williamstown, 2011: 282–289

[23] Collins M. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Philadelphia: Association for Computational Linguistics, 2002: 1–8

[24] Li Zhenghua, Zhang Min, Che Wanxiang, et al. A separately passive-aggressive training algorithm for joint pos tagging and dependency parsing // COLING. Mumbai, 2012: 1681–1698