

文章编号:1007-5321(2017)02-0016-05

DOI:10.13190/j.jbupt.2017.02.003

# 针对中国学生英文文章的词性标注方法

谭咏梅, 杨 林, 胡 单

(北京邮电大学 智能科学与技术中心, 北京 100876)

**摘要:** 提出了一种基于词向量的两层词性标注方法,使用少量人工提取的特征,大部分特征可使用词向量和第 1 层标注向量自动训练得到. 该方法将标注集分成两类,分别作为不同层的标注集. 首先,对容易标注的类别进行标注;然后,对难以标注的动词或者名词进行第 2 层标注,将其标注为具体的某类动词或名词. 利用该方法对中国学生写的英语文章进行词性标注的准确率可从 95.23% 提高到 95.63%,超过了现有基于词向量词性标注器对相同语料词性标注的准确率.

**关 键 词:** 词性标注; 中国学生; 文章; 词向量

**中图分类号:** TN911.22

**文献标志码:** A

## A Part-of-Speech Tagging Algorithm for Essay Written by Chinese English Learner

TAN Yong-mei, YANG Lin, HU Dan

(Intelligence Science and Technology Center, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract:** A tagging algorithm about two layers part-of-speech base on word embedding was proposed. Only a few artificial features are needed in this algorithm, most features are replaced by word embedding and tagging vector that is got in the first layer. In addition, the tag set is divided into two categories, which are the tag sets of different layers. The ones which are easily to be tagged are tagged firstly in the first layer. Those tags which are hardly to be tagged as noun and verb are tagged in the second layer. Using this algorithm, the accuracy of part-of-speech tagging of essays written by Chinese English learner is improved from 95.23% to 95.63%, which outperforms the state-of-art word results of part-of-speech tagging of essays written by Chinese English learner based on vector based on word embedding.

**Key words:** part-of-speech tagging; Chinese English learner; essays; word vector

词性标注是自然语言处理的一个基本任务,研究者对其做了许多有益的工作. 传统词性标注方法使用最大熵<sup>[1]</sup>、支持向量机<sup>[2]</sup>、guided learning<sup>[3]</sup>、隐马<sup>[4]</sup>等模型进行标注,这些模型大多需要进行人工特征提取.

中国学生写的英语文章常出现泊来词、专有名词、臆造词,常伴随着各种错误<sup>[5]</sup>. 由于上述问题的存在,对中国学生写的英语文章进行词性标注时,特

征提取变得十分困难. 笔者提出了两层标注的词性标注方法.

笔者将 penn TreeBank 词性标注集中的 NN、NNS、NNP、NNPS 全部用 N 代替,VB、VBD、VBG、VBN、VBP、VBZ 全部用 V 代替. 将替代后的词性标注集称为标注集  $T_1$ ,将 NN 和 VB 等分别称为标注集  $T_2(1)$  和标注集  $T_2(2)$ . 首先使用标注集  $T_1$  对英文句子进行词性标注,完成标注后,利用标注结果对

标注为 N 或者 V 的单词进行第 2 层标注。

## 1 相关工作

2003 年, Bengio 等<sup>[6]</sup>提出了一种基于词向量的语言模型, 该语言模型使用的词向量和网络参数是在无监督训练过程中得到的。考虑单词的结构, Alexandrescu 等<sup>[7]</sup>提出了一种使用单词各种自身属性的方式来表示单词的方法。2011 年, Collobert<sup>[8]</sup>提出了一种基于词向量的分类方法, 为了避免大量的特征提取工作, 使用一组词向量代替人工提取的特征。2013 年, Mikolov 等<sup>[9]</sup>提出了大型数据集中训练词向量的两种常用结构, 即 continuous bag-of-words 和 skip-gram。基于 Collobert 提出的方法, Santos 等<sup>[10]</sup>在 2014 年提出了基于字母向量表示的词性标注结构模型。该方法对英文词性标注结果的准确率为 97.32%。

## 2 两层词性标注方法

基于词向量的两层词性标注过程如图 1 所示, 在  $T_1$  上对待标句子进行词性标注, 得到第 1 层结果后, 利用该结果得到每个单词邻近词标注信息, 对标注为“N”和“V”的词进行第 2 层标注, 得到最终标注结果。

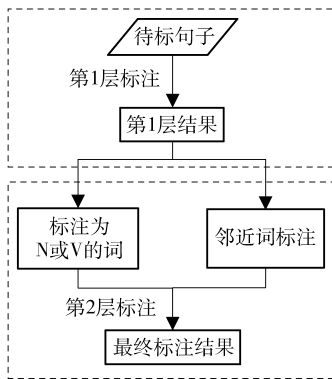


图1 基于词向量的两层词性标注方法框架

在图 1 中, 对句子进行第 1 层标注和第 2 层标注过程是基于 Collobert 提出的网络结构而改进的。其结构如图 2 所示, 在句子第 1 层标注阶段, 输入一个英文句子, 结合词向量映射表, 对句子中的每个单词输出一个概率向量, 该向量表示目标单词标注为不同词性标注  $t \in T_1$  的概率。在本结构中, 以目标单词为中心, 固定窗口大小的词向量串联作为神经网络的输入, 经过一个 3 层神经网络得到目标单词的词性概率向量。通过该结构得到每个单词在  $T_1$  标

注集的标注概率, 并结合一个标注转移概率, 使用维特比算法, 对整个句子进行词性标注。在得到第 1 层标注结果后, 结合标注信息, 对标注为“N”和“V”的词进行第 2 层标注, 从而得到最终标注结果。

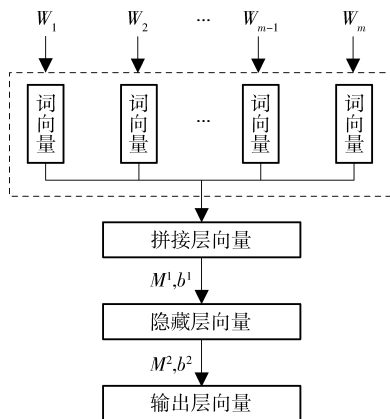


图2 词性标注网络结构

### 2.1 单词转换词向量

在第 1 层标注阶段, 本结构和 Collobert 提出的网络结构一样, 将英文句子中的每个单词转换为一个含有特征信息的词向量。例如, 词向量映射表为  $W$ , 输入一个含有  $N$  个单词的英文句子  $\{s_1, s_2, \dots, s_N\}$ , 每个单词  $s_n$  可以在该表中找到一个对应的词向量  $u_n = (a_1, a_2, \dots, a_m)$ , 其中  $m$  为词向量维度。词到词向量的转换公式为

$$u_m^N = W^w V_w^N \quad (1)$$

其中:  $w$  为词向量映射表  $W$  的大小;  $u_m^N$  由  $N$  个  $m$  维的词向量组成;  $V_w^N$  为  $N$  个单词在词向量映射表  $W$  中对应的位置信息矩阵, 在该矩阵中, 第  $i$  列向量  $v^i$  表示  $u_m^N$  中和第  $i$  个单词位置相关的向量;  $W \in \mathbf{R}^{w \times m}$  为词向量映射表,  $m$  为词向量维度。

本结构增加 3 个特征表示单词的首字母大小写情况、单词的倒数第 1 和第 2 个字母。

### 2.2 标注转换标注向量

在第 2 层标注阶段, 词性标注网络在 Collobert 提出的网络结构基础上加入了一个标注向量表。神经网络的输入层不只是词向量以及字母特征向量串联, 还包括标注向量。在此阶段, 本结构将目标词及其邻近词的标注转换为相应的标注向量。标注转换为标注向量过程和词转换为词向量类似, 其过程如式(2)所示。

$$u_{m_1}^N = W^r v_r^N \quad (2)$$

其中:  $r$  为标注向量表的大小,  $v_r^N$  为位置向量组成的

矩阵,  $\mathbf{W}^r \in \mathbf{R}^{w^r}$  为不同标注对应标注向量组成的向量表。

### 2.3 神经网络结构及输入

在第1层标注阶段,输入一个英文句子,输出英文句子的标注串。在选择合适标注串前,需求得句子中每个单词对  $T_1$  中各个标注的概率向量。这个标注概率向量的求得是基于一个自然语言处理常见假设,一个单词的词性标注依赖于该单词以及其周边单词。对于第  $t$  个单词,输入层为

$$\mathbf{l}_1 = (\mathbf{u}^{t-\frac{d}{2}} \mathbf{u}^t \mathbf{u}^{t+\frac{d}{2}})^T \quad (3)$$

其中:  $\mathbf{l}_1$  为输入层,  $d$  为窗口大小。另外,本结构使用“111”作为第1个单词前左邻近词和最后一个单词的右邻近词。接着将输入层  $\mathbf{l}_1$  输入到一个 BP 神经网络中。输出层计算如下:

$$\mathbf{o} = M^2 f(M^1 \mathbf{l}_1 + b^1) + b^2 \quad (4)$$

其中:  $M^1 \in \mathbf{R}^{h \times d \times (m+3)}$ ,  $M^2 \in \mathbf{R}^{h \times |T_1|}$ ,  $b^1 \in \mathbf{R}^h$ ,  $b^2 \in \mathbf{R}^{|T_1|}$ ,  $h$  为隐藏层数目,  $|T_1|$  为标注集大小。

本结构最后判定句子词性标注串时,借助一个转移矩阵  $\mathbf{A} \in \mathbf{R}^{|T_1| \times |T_1|}$ , 其中  $A_{ij}$  表示从词性  $i \in T_1$  到  $j \in T_1$  的转移概率。假设输入一个英文句子  $[s]_1^N = \{s_1, s_2, \dots, s_N\}$ , 利用

$$S([s]_1^N, [t]_1^N, \theta) = \sum_{n=1}^N (A_{t_{n-1}, t_n} + o_{t_n}) \quad (5)$$

其中  $\theta$  为该结构中所有的训练参数。求得该输入句子标注结果为  $[t]_1^N = \{t_1, t_2, \dots, t_N\}$  的概率值。通过比较输入句子对应不同标注串的概率值,选择最大概率值对应的标注串作为该句子的词性标注结果。在第2层标注阶段中,将第1层标注阶段得到的目标词和它邻近词的标注,根据式(2)生成标注向量,将其和词向量串联作为第2层网络输入。该阶段生成输入层为

$$\mathbf{l}_1 = (\mathbf{u}^{t-\frac{d}{2}} \mathbf{u}^t \mathbf{u}^{t+\frac{d}{2}}, \mathbf{u}^{t-\frac{g}{2}} \mathbf{u}^{t-1} \mathbf{u}^{t+1} \mathbf{u}^{t+\frac{g}{2}})^T \quad (6)$$

其中:  $\mathbf{u}^i$  为位置  $i$  处单词在第1层标注结果对应的标注向量,该向量由式(2)得到;  $g$  为目标词窗口大小。在此阶段,由式(6)得到输入层  $\mathbf{l}_1$ , 接着将其输入至一个 BP 神经网络结构中。

### 2.4 模型训练

本模型训练过程包括有监督训练和无监督训练两部分。在无监督训练过程,使用中国学生写的英文文章作为无标语料,训练得到一个词向量映射表。在有监督训练过程中,将无监督训练得到的词向量映射表作为词性标注模型中  $\mathbf{W}_1$  的初始化值,然后

再利用有标训练语料,对第1层神经网络结构进行有监督训练。

#### 2.4.1 无监督训练过程

在本结构中,词向量充当单词特征的角色,并且 Collobert 实验证明,使用大量无标注语料对其进行无监督训练可以使结果更好<sup>[8]</sup>。无监督训练词向量映射表的方法有多种: Bigram 神经网络用于生成词对应的词向量<sup>[11]</sup>、训练一个基于神经网络的语言模型得到词向量<sup>[6]</sup>、使用 CBOW (continuous bag-of-words model) 和 Skip-gram 模型训练词向量<sup>[9]</sup>。

#### 2.4.2 有监督训练过程

以第1层神经网络训练过程为例,首先通过在有标注训练集合  $D$  上,使用极大似然目标函数值和梯度上升算法更新模型参数。极大似然目标函数为

$$\mathbf{W}, \theta_1 \rightarrow \sum_{([s]_1^N, [t]_1^N, \theta_2) \in D} \lg P([t]_1^N | [s]_1^N, \theta_1, \mathbf{W}) \quad (7)$$

其中:  $[s]_1^N$  表示训练集中的英文句子,  $[t]_1^N$  表示英文句子对应的词性标注串。定义:

$$\lg P([t]_1^N | [s]_1^N, \theta_1, \mathbf{W}) = S([s]_1^N, [t]_1^N, \theta_1, \mathbf{W}) - \lg(\sum_{[j]_1^N} e^{S([s]_1^N, [j]_1^N, \theta_1, \mathbf{W})}) \quad (8)$$

其中  $\forall [j]_1^N$  表示任意长度为  $N$  的词性标注串。

本结构采用梯度上升的方法,通过式(7),对函数中的参数值进行更新。训练时,每次迭代过程对参数  $\theta_1$  更新如下:

$$\mathbf{W}, \theta_1 < -\theta_1, \mathbf{W} + \lambda \frac{\partial \lg P([t]_1^N | [s]_1^N, \theta_1, \mathbf{W})}{\partial \theta_1, \mathbf{W}} \quad (9)$$

其中:  $\lambda$  为模型参数更新的变化率;  $\theta_1$  为第1层神经网络参数,包括  $M^1$ 、 $b^1$  和转移矩阵  $\mathbf{A}$ 。

## 3 实验

### 3.1 实验数据

本实验的有标注语料和测试语料来自于吴坤的词性标注实验语料<sup>[12]</sup>。该语料也来自批改网收集的中国学生写的英语文章,选取其中 9 077 条作为有监督训练语料,1 346 条作为测试语料。该语料中常有泊来词、专有名词、臆造词等非常见词,同时伴随着拼写错误,单复数误用,时态或语态错误,冠词、介词、代词、非谓语动词的误用等现象。针对上述问题,使用的词向量由学生作文语料训练得到,一定程度上,对含有语法错误的句子进行了拟合。另外,提取少量的人工特征,用来弥补因拼写错误、未登录词等造成的词性标注不准确的问题。

3.2 实验超参数

实验超参数如表 1 所示.

表 1 实验超参数表

参数	意义	值
$d$	第 1 层标注网络窗口大小	5
$g$	第 2 层标注网络窗口大小	7
$m$	词向量维度	50
$h$	标注网络隐藏层数	350
$\lambda$	参数变化率	0.001
$w$	粗向量词表大小	179 113
$r$	第 1 层标注集大小	37

3.3 主要实验

面向中国学生写的英文文章的词性标注是基于词向量的词性标注,在这个标注过程中,使用词向量代替大部分的人工提取的特征. 通过对比不同方法初始化词向量映射表得到的词性标注准确率,选择一种合适的无监督训练词向量的方法. 笔者提出的基于词向量的词性标注是双层的标注结构,对于双层标注的有效性可通过下面的实验证明.

实验根据第 2 节设计的两层标注方法,首先使用学生作文进行无监督训练,得到词向量表;然后使用该词向量表,对有监督的训练中的单词进行初始化. 在实验中,为验证两层方法的有效性,对不同的无监督训练词向量方法、标注方法(是否双层标注)进行组合并分别进行实验. 实验结果如表 2 所示,其中,S1、S2、S3、S4、S5 分别表示随机值初始化词向量、Bigram 神经网络生成词向量、训练一个语言模型生成词向量、使用 CBOW 模型生成词向量、使用 Skip-gram 模型生成词向量等 5 种不同词向量生成方式;C1 表示只使用第 1 层标注方法进行词性标注,C2 表示只使用第 2 层标注方法进行词性标注,C3 表示基于词向量的两层词性标注结构.

表 2 不同实验组合实验结果

组合	准确率/%	组合	准确率/%	组合	准确率/%
S1 + C1	75.43	S2 + C3	85.63	S4 + C2	92.96
S1 + C2	73.21	S3 + C1	94.69	S4 + C3	93.62
S1 + C3	74.77	S3 + C2	93.20	S5 + C1	95.31
S2 + C1	84.86	S3 + C3	94.68	S5 + C2	94.56
S2 + C2	84.27	S4 + C1	92.10	S5 + C3	95.63

如表 2 所示,两层的标注方法与一层的标注方法相比,词性标注的准确率较高,可以证明两层结构

的有效性;使用 CBOW 或者 Skip-gram 来训练词向量用于标注模型相比训练语言模型生成词向量或者随机设置的方式用于标注模型,得到的标注准确率更高,证明了所提出的基于词向量的标注模型的有效性;当使用 Skip-gram 模型初始化词向量映射表,并且使用两层标注方法对中国学生写的英文文章词性标注时,词性标注的准确率最高.

标注器 GENIA Tagger 和 Senna 在词性标注中,都有较好的效果. 通过实验比较了这两种标注器和所提出的两层标注模型 S 在中国学生写的英文文章上的词性标注准确率,实验结果如表 3 所示. GENIA Tagger 和 Senna 在华尔街日报语料中词性标注准确率分别达到了 94.49% 和 95.23%,所提出的词性标注方法相对标注器 GENIA Tagger 和 Senna 都有所提高. 由表 3 可知,虽然 Senna 和 S 只使用了少量的人工特征,但是这两种标注效果并不比 GENIA Tagger 差. 这说明在标注过程中,词向量确实可以代替部分人工提取的特征. 除此之外,由表 3 可以看出,S 的标注准确率比 Senna 高 0.4%. 这说明所提出的方法比 Senna 更适用于对中国学生写的英文文章进行词性标注.

表 3 不同模型的标注准确率 %

标注器	中国学生写作语料	华尔街日报语料
GENIA Tagger	94.49	95.87
Senna	95.23	95.14
S	95.63	93.90

为了测试标注集对其他语料的适用性,使用华尔街日报 23 章语料作为测试集,利用 GENIA Tagger、senna 和所提模型对其标注. 标注结果如表 3 所示. 基于词向量的两层词性标注方法对华尔街日报语料标注准确率只有 93.90%,低于其他两种标注器的标注准确率,原因可能是基于词向量的两层词性标注方法受训练语料的影响较大.

4 结束语

提出了一种基于词向量针对中国学生写的英语文章的词性标注方法. 该方法首先对输入的句子进行第 1 层标注,将所有的名词和动词分别使用“N”和“V”进行标注,其他词性的单词进行直接标注;然后利用第 1 层标注结果对不易标注的词(第 1 层中标注为“N”和“V”的单词)进行第 2 层标注,将不易标注的词进行更加详细的标注.



笔者提出的标注结构是一种基于词向量的神经网络,利用该神经网络对中国学生写的英语文章进行词性标注,标注准确率从 95.23% 提高到 95.63%,超过了现有的词性标注器对相同语料词性标注的准确率。

## 参考文献:

- [1] Toutanova K, Manning C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger [C] // Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora; Held in Conjunction with the, Meeting of the Association for Computational Linguistics. Hong Kong: Association for Computational Linguistics, 2000: 63-70.
- [2] Màrquez L, Giménez J. A general pos tagger generator based on support vector machines[J]. JMLR, 2004(5): 1253-1286.
- [3] Shen L, Satta G, Joshi A. Guided learning for bidirectional sequence classification[C] // Meeting of the Association for Computational Linguistics, Prague, Czech Republic: Association for Computational Linguistics. 2007: 760-767.
- [4] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, et al. Improved part-of-speech tagging for online conversational text with word clusters[C] // Proceedings of NAACLHLT 2013. Los Angeles, California: Association for Computational Linguistics, 2013: 380-390.
- [5] 李红, 大学生英语写作常见错误归类分析[J]. 当代教育论坛: 学科教育研究, 2006(8): 120-121.  
Li Hong. The common errors analysis of college english-writing[J]. Forum on Contemporary Education, 2006(8): 120-121.
- [6] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
- [7] Alexandrescu Andrei, Kirchhoff Katrin. Factored neural language models [C] // Proceedings of the Human Language Technology Conference of the NAACL. New York City, USA: [s. n. ], 2006: 1-4.
- [8] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.
- [9] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [EB/OL]. 2013-09-05. <https://arxiv.org/abs/1301.3781>.
- [10] Santos C N D, Zadrozny B. Learning character-level representations for part-of-speech tagging[C] // International Conference on Machine Learning. Beijing: ICML, 2014: 1818-1826.
- [11] Collobert R. Deep learning for efficient discriminative parsing[C] // The 14<sup>th</sup> International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, Florida: AISTATS, 2011: 224-232.
- [12] 谭咏梅, 吴坤. 面向英语文章的词性标注算法[J]. 北京邮电大学学报, 2014, 37(6): 120-124.  
Tan Yongmei, Wu Kun. A part-of-speech tagging method for English essay[J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37(6): 120-124.