基于最大熵模型的冠词错误纠正系统①

陈朝才[1], 吴 敏[1], 吴桂兴[2], 郭 燕[2]

[1](中国科学技术大学 现代教育技术中心, 合肥 230026)
[2](中国科学技术大学 苏州研究院, 苏州 235123)

摘 要: 研究了英语语法中冠词错误的计算机自动纠正. 首先对冠词使用的错误进行定义分类, 并考虑到可能出现冠词缺失的情况, 通过采用基于最大熵模型的分类器, 选择包含上下文、上下文词性、短语结构等特征, 在训练集上进行模型预的训练, 然后使用模型对于输入句子进行预测并纠正存在的使用错误. 在 NUCLE 语料的实验中, 给出了语料处理、模型特点、训练语料的大小对于测试集效果的影响, 并且比较了自然语言处理中非常通用的朴素贝叶斯模型的结果, 还根据英语语法中存在的错误特点对模型进行改进, 最后在测试数据达到 35.48%的 F 值, 相较于 CoNLL2013 的 shared task 中最好结果有小幅提升.

关键词: 冠词错误; 计算机自动纠正; 最大熵模型

Article Error Correction System Based on Maximum Entropy Model

CHEN Zhao-Cai[1], WU Min[1], WU Gui-Xing[2], GUO Yan[2]

[1](Center of Modern Educational Technology, University of Science and Technology of China, Hefei 230026, China)
[2](Suzhou Institute of University of Science and Technology of China, Suzhou 235123, China)

Abstract: Computer automation correction of article errors in English grammar is been studied. First we define the categories of article errors, and missing articles is also included, by using a maximum entropy model, extracting features covering context, part of speech, noun phrase structure and so on, training the model on the training corpus, then use the model to predict and correct the article errors of an input sentence. In the experiment on NUCLE corpus, effects of corpus preprocess, model types and the size of the training corpus are discussed. We make a comparison with the popular Naive Bayes model, at last we introduce the characters of English grammar to improve the model, a F-score of 35.48% is achieved, the result is slightly better than the best result in CoNLL 2013 shared task.

Key words: aticle errors; computer automation correction; maximum entropy model

冠词在英语写作中是非常普遍的一种限定词, 虽然冠词只包含“a/an”和“the”两种, 却是英语学习者的一个难点, 引起了很多研究者的注意. 随着自然语言处理和机器学习的发展, 越来越多的研究尝试使用算法让计算机来自动解决此类问题, 本文尝试采用最大熵模型, 通过提取上下文特征, 将冠词的使用当做分类问题处理, 从而对英语写作中的冠词错误进行纠正.

本文组织结构如下, 第一节介绍冠词错误纠正的相关研究, 第二节介绍最大熵模型及处理过程, 第三节介绍系统模块以及系统整个流程, 第四节给出在测试集语料上的实验结果并进行相关分析, 最后进行总结和展望.

1 相关研究

在最近的研究中, 使用计算机辅助帮助英文写作得到了广泛关注, 而在写作中, 如何帮助发现作文中可能出现的语法错误是一大难题. 在语法错误中, 存在主谓一致、名词单复数、冠词使用错误等类型, 在这其中, 冠词使用错误占比很高. 纠正冠词使用错误的方法有很多种, 主要分为两类, 一类基于规则, 可以

① 收稿时间:2014-12-12;收到修改稿时间:2015-02-02

尝试运用大语料库中统计的 n 元模型来做检测，如在 CoNLL 2013 的 shared task 中，Ting-hui Kao, Yu-Wei Chang 等[1]利用当前冠词的一个 4 元窗口，通过查找 google web 1T 语料库，综合考虑 n-gram 出现频率、PMI 值等因素，设置对应阀值来判断当前冠词是否使用错误. 另一类是基于统计来进行纠错，Golding, Roth 等[2]使用基于 Winnow 算法的分类器来纠正上下文相关的拼写错误，选出可能混淆的单词组合构成混淆集，从混淆词出现的上下文提取出特征向量，通过训练，分类器可以选择在当前上下文中可能性最大的单词. 在本文中采取的方法类似，冠词对应的混淆集即为{a/an, the, NULL}，模型根据提取出的特征向量选出概率最大的混淆次作为系统纠正.

Alla Rozovskaya, Kai-Wei Chang 等[3]使用基于平均感知器[4]和 Naïve Bayes 的组合来对介词错误进行纠错，他们将最常见的 10 个介词作为混淆集，运用词的 n-gram 特征，进行训练，通过上下文来判断和纠正一部分介词错误，例如在句子 I want to go home .中通过上下文特征，可能发现上文中的 want 会经常和 to 进行组合，那么通过对应的统计即可判断下文的介词为 to，而非其他介词.

2 基于最大熵模型的冠词纠错

2.1 最大熵模型基本原理

最大熵模型[5]的基本思想是在满足所有已知因素的情况下，使得熵达到最大，也就是将所有的未知因素排除在外. 在我们的问题中，对于一系列影响冠词使用判断的因素，表示为一个特征向量 X，那么我们就要找到使得 p(冠词|X)达到最大的那个冠词，在最大熵模型中要求 p(冠词|X)满足一定约束条件的情况下，使得下式定义的熵取得最大值:

$$p = \arg\max_p(H(p)) = \arg\max_p(-p(x)\log_2 p(x)$$

约束条件可以通过如下方式来描述

$$f_i(a,b) = \begin{cases} 1, (a = inde) \wedge (b = cat) \\ 0, otherwise \end{cases}$$

i = 1,2,3,…,n, f_i 是最大熵模型的特征，n 为所有的特征数，这些特征描述了向量 X 与冠词选择之间的关系. 概率 p(冠词|X)必须满足这些特征的约束，定义一个受限的分布为:

$$E_p f_i = E_{\tilde{p}} f_i$$

其中:

$$E_{\tilde{p}} f_i = \sum_{a,b} \tilde{p}(a,b) f_i(a,b)$$

$$E_p f_i = \sum_{a,b} \tilde{p}(b) p(a \mid b) f_i(a,b)$$

$\tilde{p}(b)$ 和 $\tilde{p}(a,b)$ 是训练集中观察到的经验分布，那么问题就是从受限的分布中找到一个熵达到最大的分布，可以求出解为

$$p^*(a \mid b) = \frac{1}{\pi(b)} \exp(\sum_{i=1}^{n} \lambda_i f_i(a,b))$$

其中，$\pi(b)$ 是归一化因子.

$$\pi(b) = \exp(\sum_{i=1}^{n} \lambda_i f_i(a,b))$$

2.2 问题定义及模型的建立

我们将在英语作文中出现的冠词错误分为三种，①冠词多余，如 the air cargo of the/NULL valujet plane was on fire after the plane had taken off .中这里的 the 多余应删掉. ②冠词缺失，如 in NULL/the modern digital world , electronic products are widely used .中 modern digital world 前面缺少冠词，应加上 the. ③冠词选择错误，如 Here is a/the man .这里 man 前面为特指，应为 the. 那么我们可以将对冠词的判断作为一个分类问题处理. 在本文中，我们定义三类分别为 INDE、DEFI、NULL，其中 INDE 指不定冠词 a 或者 an(由于 a/an 的判断是基于规则的，故可以当做一类)，DEFI 指定冠词 the，NULL 指无冠词,那么目标就是找到一个$\mathbf{y \in \{INDE, DEFI, NULL\}}$，使得 p(y|X)达到最大，在最大熵模型中通过求解方程使得熵达到最大值. 另外，由于冠词只会出现在名词或名词词组前，故在本文中将所有被标注为名词或名词词组前都作为可能出现冠词的位置，如 in modern digital world, modern digital world 被标记为名词短语，因此假设该名词词组前可能会出现冠词，也提取出对应特征进行判断.

2.3 处理过程

(1) 对输入句子使用 stanford parser[6]进行词性标注和语块标注，如句子 all passengers and pilots were died .被标注为(ROOT (S (NP (DT all) (NNS passengers) (CC and) (NNS pilots)) (VP (VBD were) (VP (VBN died))) (. .))), 在这个句子中, passengers 被标注为名词，died 被标注为动词，all passengers and pilots 被标注为名词短语, were died 被标注为动词短语.

(2) 找到输入句子存在冠词(不定冠词或定冠词)和可能缺失冠词的位置，在英语中可能出现冠词的位置只有在名词短语(被语块标注为 NP)之前，如在上句

中, 出现的名词短语为 all passengers and pilots, 那么认为该短语前可能存在冠词.

(3) 对 2 中找到位置提取出表中对应的特征, 且对应的特征标签为当原位置是不定冠词时为 INDE, 定冠词时为 DEFI, 没有冠词时为 NULL, 如在上句中, all passengers and pilots 这个名词短语之前未出现冠词, 那么标签应为 NULL.

(4) 在训练集中进行最大熵模型的训练. 对于测试集, 使用最大熵模型进行预测, 若预测结果与原结果一致, 则不进行纠正, 若预测结果与原结果不同, 则纠正模型预测结果, 如在 all passengers and pilots 中, 若模型预测为 DEFI, 那么在该短语之前添加冠词 the.

3 冠词纠错系统设计

3.1 系统模块

(1) 语料预处理: 对每一个句子, 将其转为小写, 并检查是否为合法输入.

(2) 词性以及语块标注模块: 对训练集或输入进行词性标注和语块标注, 为此我们封装了一个 stanford parser 的接口.

(3) 特征提取模块: 特征是模型的主要部分, 我们的特征中一共包括 8 个方面, 其中主要包含 N-gram 特征、词性 N-gram 特征、名词短语特征等, 具体特征见表 1. 本文模型中用到的特征, wB、wA、pB、pA 分别表示当前词之前的词、之后的词、之前词的词性以及之后词的词性. headWord 表示名词短语中的第一个词(非冠词), lastWord 表示名词短语中最后一个词, NC 表示在名词短语中从前往后第一个被标记为名词的词, NNS 表示名词短语中是否有复数形式的名词, Verb 表示当名词短语之后紧跟动词时该动词, Preposition 表示当名词短语之前为介词时该介词, npWords 和 npTags 表示名词短语中所有的词及所有词的词性.

表 1 具体特征表

特征类型	特征描述
Word n-gram	wB, w2B, w3B, wA, w2A, w3A, wBwA, w2BwB, wAw2A, w3Bw2BwB, w2BwBwA, wBwAw2A, wAw2Aw3A, w4Bw3Bw2BwB, w3Bw2BwBwA, w2BwBwAw2A, wBwAw2Aw3A, wAw2Aw3Aw4A
POS n-gram	pB, p2B, p3B, pA, p2A, p3A, pBpA, p2BpB, pAp2A, pBwB, pAwA, p2Bw2B, p2Aw2A, p2BpBpA, pBpAp2A, pAp2Ap3A
NP1	headWord, lastWord, npWords, NC, NNS, adj&headWord, adjTag&headWord, adj&NC, adjTag&NC, npTags&headWord, npTags&NC
NP2	headWord&headPOS
WordsAfterNP	headWord&wordAfterNP, npWords&wordAfterNP, headWord&2wordsAfterNP, npWords&2wordsAfterNP, headWord&3wordsAfterNP, npWords&3wordsAfterNP
WordsBeforeNP	wB&f_i i ∈ NP1
Verb	Verb, verb&f_i i ∈ NP1
Prepositions	Prep&f_i i ∈ NP1

(4) 模型训练模块: 在训练集上进行模型的训练, 这里封装一个最大熵模型的实现.

(5) 系统纠正模块: 根据模型预测结果对原句子进行纠正.系统实现之后对于任意输入的英语句子, 就可以检查其中存在的冠词使用错误, 若使用正确, 则不纠正, 若使用错误, 则根据系统输出进行纠正后输出.

3.2 系统结构流程图

在图 1 中表述了整个英语冠词错误纠正系统的检查过程, 首先通过训练集提取特征进行模型的训练, 然后对于每条输入的句子, 我们先进行词性标注和语块标注, 然后提取出对应的特征, 特征输入到模型进行预测, 若预测结果和原标签相同, 则不进行纠正, 若不同, 则纠正为系统预测结果并输出.

4 实验结果及分析

4.1 语料介绍及预处理

实验采用 NUCLE-release2.2 语料[7]，该语料为此部分新加坡国立大学收集的作文并进行人工标注，在语料中一共包含 28 种错误类型，其中 ArtOrDet 表示限

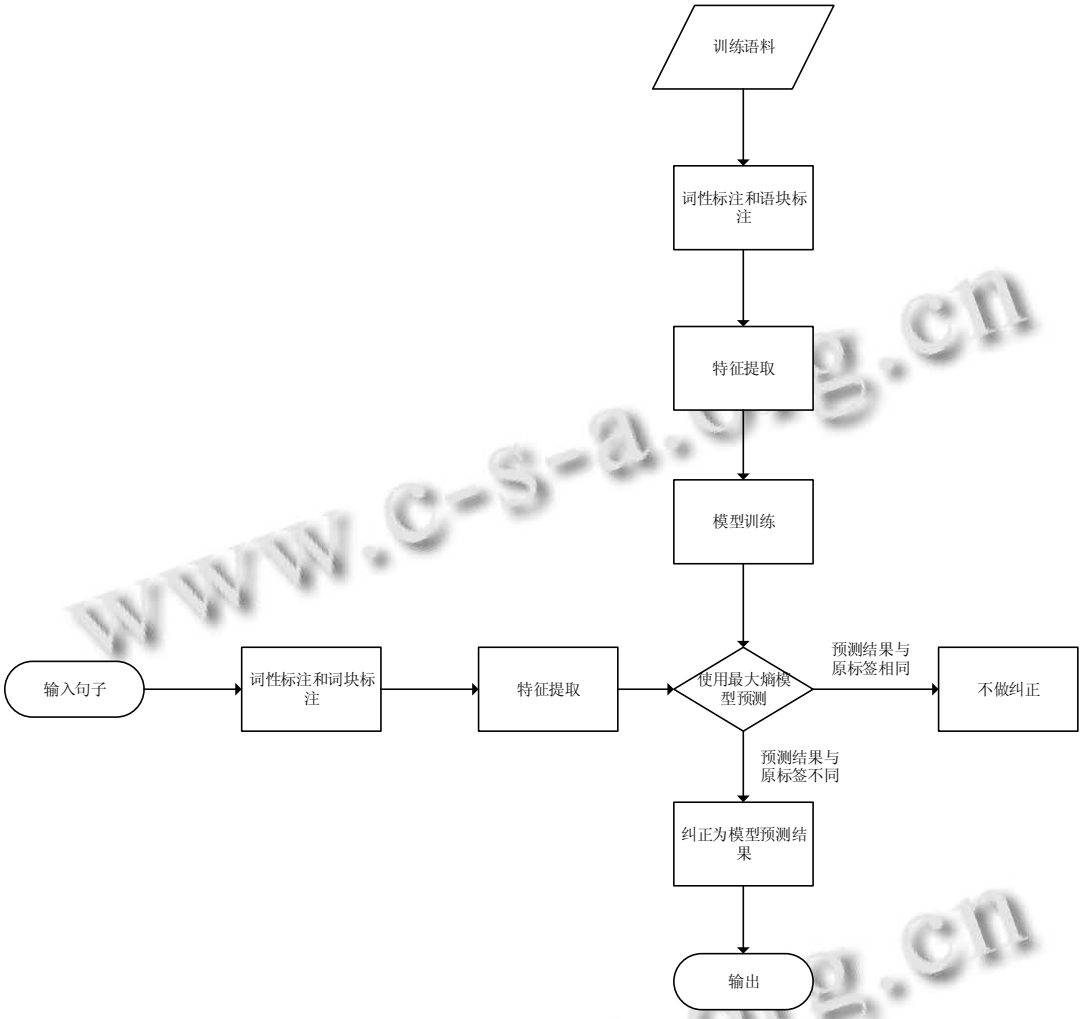

图 1　英语冠词错误纠正系统流程图

定词错误，本文中只关注限定词中的冠词类错误. 训练集及测试集中详细数据见表 2.

表 2　训练集及测试集中句子总数, token 总数，限定词错误数以及限定词错误占总错误比例

	训练集	测试集
Sentences	57151	1381
Tokens	1161567	29207
ArtOrDet 错误数及占比	6658/14.9%	690/19.9%

为方便实验，训练集及测试集中所有句子均被转至小写，训练集中所有错误被纠正以得到正确训练语料且去除包含括号和超链接的句子(这些句子会影响 parser 的准确性，4.2 将给出对比结果)，测试集中错误不进行纠正.

4.2 实验结果

实验的评价标准采用正确率、召回率以及 F 值.

在预处理中，我们去除了所有包含括号和超链接的句子，表 3 给出处理前后的结果对比以及与 CoNLL 2013 shared task[8]中最好结果的对比.

表 3　对语料进行预处理前后的测试集结果对比

	Precision	Recall	F
预处理前	23.78	55.80	33.35
预处理后	24.42	56.38	34.08
Shared task	25.65	47.84	33.40

可见去除掉这些会让词性标注不准确的句子后，并没有使结果变差，反而有一定的提升，而且去除掉这些句子后，模型的训练速度也将加快，效率更高，

另外通过比较看出，我们的系统相较于 CoNLL 2013 shared task 中的最好结果有小幅度提升.

在最大熵模型中，不同优化算法以及迭代次数会对结果产生一定影响，表 4 给出迭代次数与 F 值的关系，从表 4 中可以看出随着迭代次数的增加，F 值会有小幅度增加，但并无很明显提升，而且随着迭代次数的增加，训练时间会大幅度增加，因此后续实验中均设置迭代次数为 1.

表 4 最大熵模型迭代次数对于测试集结果的影响

迭代次数	1	2	3	4	5	6	7	8	9	10	100
F	34.08	34.08	34.11	34.12	34.18	34.18	34.28	34.26	34.26	34.24	35.48

在自然语言处理中，朴素贝叶斯模型是一种非常流行的分类模型，因此我们对比了两种模型在测试集中的效果(两种模型使用相同的特征，见表 5).

表 5 朴素贝叶斯模型和最大熵模型在测试集上的表现对比

	Precision	Recall	F
朴素贝叶斯模型	16.13	41.59	23.25
最大熵模型	24.42	56.38	34.08

从表 5 中可以看出最大熵模型的效果远远好于朴素贝叶斯模型，朴素贝叶斯的模型优点在于模型训练速度较快，且较易实现，而最大熵模型是判别模型，模型复杂度更高，虽然效果很好，但训练速度较慢.

另外，我们考虑到实验中所使用到的语料还是较少，因此另外加入一部分维基语料进行实验对比，表 6 给出添加语料后的测试集结果，可以看出加入另外的训练语料后效果有小幅度提升，但在实验中模型训练占用的内存以及训练时间将会大幅度增加.

表 6 添加维基 10w、20w 条语料后测试集效果对比

	Precision	Recall	F
NUCLE	24.42	56.38	34.08
NUCLE+维基(10w 条)	24.27	56.81	34.01
NUCLE+维基(20w 条)	24.54	57.39	34.38

英语写作中，学习者对于冠词的掌握是有偏向的，在实验中我们发现不同错误类型出现的次数具有很大差别(表 7 和表 8 分别训练集和测试集中几种错误类型出现的次数统计).

表 7 训练集中不同错误出现数，表中表示标注语料中出现的纠正数

	NULL	INDE	DEFI
NULL		1020	3007
INDE	204		93
DEFI	1814	200	

表 8 测试集中不同错误出现数，表中表示标注语料中出现的纠正数

	NULL	INDE	DEFI
NULL		82	164
INDE	38		11
DEFI	339	30	

从表 7 和表 8 中可以看出，冠词的缺失错误和定冠词的误用占比最大，在训练集中共出现 4027 处冠词缺失，而定冠词多余共出现 1814 次，在测试集中也可发现对应情况，这表明，在英语写作中，学习者犯的错误是有偏向的，对于一般的 a 或者 an 这种不定冠词的使用错误很少，而大部分错误集中在定冠词和不使用冠词的情况下. 在我们的模型中没有考虑到这种情况，因此我们尝试在模型中引入错误类型的的特征，让模型去发现这种错误类型的不平衡占比，为此我们将原来的训练样本的标签以一定概率替换为其他标签，且加入错误率来控制引入的错误数，如当前样本标签为 NULL，错误率为 0.8，那么以 0.8 的概率引入错误替换，且将其以 1814/(1814+204)的概率替换为 DEFI，即在训练集中引入错误，并将发生错误的标签作为一维特征加入模型. 在表 9 中给出引入错误类型先验后不同错误率对于结果的影响. 从表 9 中看出，错误类型特征的加入在一定的错误率下对于模型有小幅度的提升，说明引入错误类型对于模型效果有一定的效果. 作为比较，在表 10 中给出使用朴素贝叶斯模型时算法的效果，可以看出在朴素贝叶斯模型中引入错误类型特征后在错误率小于 1 时均有小幅度提升.

表 9 不同错误率在最大熵模型中对于测试集结果的影响

错误率	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
F	33.97	33.83	33.79	33.94	34.01	34.15	34.11	34.04	34.12	33.94

表 10 不同错误率在朴素贝叶斯模型中对于测试集结

错误率	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
F	22.96	23.50	23.49	23.61	23.65	23.68	23.65	23.55	23.68	23.47

4.3 结果分析

下面分别给出一些系统正确的判断和错误的判断.

正确的判断:

① 原句: however , the dark side of using surveillance technology is that it results in the invasion of privacy .

系统纠正: however , the dark side of using surveillance technology is that it results in the/NULL invasion of privacy .

② 原句: many of this surveillance are being implemented in the government sectors , military areas to enhance their security .

系统纠正: many of this surveillance are being implemented in the/NULL government sectors , military areas to enhance their security .

③原句: it is useful to the government and the people in the law sector .

系统纠正: it is useful to the government and the/NULL people in the law sector .

错误的判断:

① 原句: risks can be analyzed if there is a record of what was happened in the past .

系统纠正: risks can be analyzed if there is a/NULL record of what was happened in the past .

② 原句: the air cargo of the valujet plane was on fire after the plane had taken off .

系统输出: air cargo of the/NULL valujet plane was on fire after plane had taken off .

③ 原句: they pull up the age to retire .

系统输出: they pull up the/NULL age to retire .

从输出中可以看出, 模型倾向于做出删除名词短语前面介词的选择, 从表 11 中可以看出训练集中名词短语前为 NULL(即无冠词)的情况占了大多数, 所以在模型训练过程中, 会更倾向于做出没有介词的判断.

表 11 训练集中三种类型的样本数

	INDE	DEFI	NULL
样本数	24644	88360	171605

5 总结和展望

在本文中, 通过运用最大熵模型, 提取出可能出现冠词的上下文特征, 训练模型, 用来对作文文本中出现的冠词错误做出纠正, 并深入探讨了模型参数、语料以及特征对于模型结果的影响, 并且还引入错误先验, 最终模型取得 35.48%的 F 值. 在实验过程中, 模型还是发生很多误判, 因此应该更深入的研究基本特征对于冠词选择的影响, 以及一些基于语义上的特征, 另外冠词的一些使用还受到某些规则的约束, 可以考虑和规则系统的融合. 在后续的研究中, 还可用此类方法对于其他语法错误类型进行纠正.

参考文献

1 Kao TH, Chang YW, Chiu HW, et al. CoNLL-2013 Shared Task: Grammatical Error Correction NTHU System Description. CoNLL-2013, 2013, 20.

2 Golding AR, Roth D. A winnow-based approach to context-sensitive spelling correction. Machine Learning, 1999, 34(1-3): 107–130.

3 Rozovskaya A, Sammons M, Roth D. The UI system in the HOO 2012 shared task on error correction. Proc. of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics. 2012. 272–280.

4 Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Machine Learning, 1999, 37(3): 277–296.

5 Berger AL, Della Pietra SA, Della Pietra VJ. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1): 39–71.

6 Klein D, Manning CD. Accurate unlexicalized parsing. Proc. of the 41st Meeting of the Association for Computational Linguistics. 2003. 423–430.

7 Dahlmeier D, Ng HT, Wu SM. Building a large annotated corpus of learner English: The NUS corpus of learner English. Proc. of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. 2013. 22–31.

8 Ng HT, Wu SM, Wu Y, et al. The conll-2013 shared task on grammatical error correction. Proc. of CoNLL. 2013.