基于介词向量的英语真词错误检查算法①

霍娟娟[1,2], 吴 敏[1,2], 吴桂兴[2], 郭 燕[2], 陈朝才[1,2], 杜一民[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: 在基于 Winnow 算法的基础上引入混淆词和介词搭配的方法. 首先通过混淆集获得训练集, 对训练集进行预处理后利用文本特征提取方法获得特征词集, 然后对特征词集进行 Winnow 训练得到带有权重的特征词集并把出现在混淆词后的介词提取出来生成介词向量, 最后从测试集提取特征并进行结合 Winnow 算法和混淆词与介词搭配方法的测试得到真词错误检查的结果. 混淆词与介词搭配方法的加入使得某些混淆词的正确率、召回率以及 F1 测度提高了 10% ~ 20%, 有的甚至提高到了 100%.

关键词: 真词错误; 介词; Winnow

English Real-Word Errors Checking Algorithm Based on Preposition Vector

HUO Juan-Juan[1,2], WU Min[1,2], WU Gui-Xing[2], GUO Yan[2], CHEN Zhao-Cai[1,2], DU Yi-Min[2]

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

Abstract: This paper introduces the method of collocation of confusion words and prepositions based on Winnow algorithm. Firstly, we obtain training sets by confusion sets. After preprocessing the training sets, we use the text feature extracted method to obtain feature sets. Secondly, we get the feature sets with weights by training on the feature sets based on Winnow and extract the prepositions which appear after the confusion words to generate the preposition vectors. Finally, we extract features from the test sets and the test sets and get the real-word errors checking results by the test which combines Winnow algorithm and the method of collocation of confusion words and prepositions. The correct rate, recall rate and F1 measure of some confusion words are improved by 10% ~ 20% when we join the method of collocation of confusion words and prepositions, some even up to 100%.

Key words: real-word errors; preposition; Winnow

1 引言

随着计算机与互联网技术的发展, 计算机辅助教学和 e-Learning 已经在教育领域起到越来越重要的作用[1]. 在英语考试中, 自动化考试系统基本已经能够实现对考试中客观题目的自动评分, 这很大程度上减轻了考试的阅卷工作量. 但是在主观题方面, 尤其是英语作文, 由于其答案的复杂程度高, 涉及到的技术难度也高, 因而对其实行自动评分要复杂的多[2]. 众所周知, 英语作文的批改量大且耗时较长, 人工批改受主观性影响存在偏差. 如果我们能够实现英语作文的自动评分, 就会很大程度上解决这些问题[3].

在英语作文中我们最常见的错误就是单词拼写错误, 单词拼写错误可以分为两种类型: 非词错误和真词错误. 非词错误指的是输入的这个单词在字典里不存在, 拼写有误; 而真词错误指的是输入的这个单词在字典里存在, 和写作者原本想输入的单词很相似[4,5]. 非词错误的检错方法现在也比较成熟, 由于真词错误的检查要比非词错误的检查复杂的多, 故我们这里通过在基于 Winnow 算法的基础上引入混淆词和介词搭配的方法实现对真词错误的检查. 举个例子: “adapt”和

① 收稿时间:2014-07-02;收到修改稿时间:2014-07-30

“adopt”都是字典里存在的单词, 但是二者在拼写上十分相似, 很容易造成拼写错误, 现在有个句子—“I will adapt the advice”, 很明显句子中的谓语不应该是“adapt”, 因为“adapt”接宾语是要加介词的, 根据检查、分析得知这里的谓语应该是“adopt”, 这里就犯了所谓的真词错误. 接下来我们介绍的真词错误的检查就是针对这种类型错误.

2 Winnow算法简介

Winnow 算法在应用于英语真词检错时, 是通过引入混淆词来实现的. Winnow 分类算法实际上是一种线性分类算法, 算法中的每个数据对应两个值, 一个值表示某混淆词对应的特征单词是否出现, 用 X 表示, 另一个是 X 所对应的权重 W. 我们通过下面介绍的初始化过程、训练过程和测试过程来描述 Winnow 算法如何调整该算法中的两个参数 α、β 以达到正确分类文本的目的, 算法中的 θ 参数是根据实验结果而定的阈值参数.

(1)初始化过程: 我们利用向量空间模型来表示文本, 对于某个混淆词, 其周围的 n 个特征单词(这里的特征单词是指在训练集中提取的混淆词周围出现的次数较多的单词, 我们可以设定一个阈值, 如果某一混淆词周围出现的某个单词的次数大于这个阈值, 我们就将这个单词提取出来作为这个混淆词的特征单词)可以用一维向量 $X=(x_1, x_2, ..., x_n)$(x_i 为 1 表示该特征出现, 为 0 表示该特征没有出现)来表示, 对应于特征单词的 n 个权重可以用一维向量 $W=(w_1, w_2, ..., w_n)$来表示[6], 如果

$$\sum_{i=1}^{n} w_i x_i > \theta(\theta \text{ 是设定的阈值})$$

则判定为 1, 表示这个单词正确出现, 即分类正确; 如果

$$\sum_{i=1}^{n} w_i x_i < \theta(\theta \text{ 是设定的阈值})$$

表示分类错误, 此时我们需要调整权重向量 ω_i 来减小分类错误.

(2)训练过程: 这里主要是对分类错误的训练文本进行权重的更新[7,8]:

① 如果 $\sum_{j=1}^{d} w_j x_j > \theta$, 则表示当前的分类器判断训练的文本是属于该类的, 要是实际上训练的文本并不属于该类, 那么就要降低对应的权重: 对于 j=1,2,...,d, 若 x_j 不等于 0, 那么 $w_j = \beta w_j$ ($0<\beta<1$);

② 如果 $\sum_{j=1}^{d} w_j x_j < \theta$, 则表示当前的分类器判断训练的文本并不是属于该类的, 要是实际上训练的文本确实属于该类, 那么就要提高对应的权重: 对于 j=1,2,...,d, 若 x_j 不等于 0, 那么 $w_j = \alpha w_j$ ($\alpha>1$).

(3)测试过程: 经过前面两个过程, Winnow 分类器已经构造完成, 接下来只需要在测试集上进行测试即可, 看分类器能否找出真词错误并进行纠正, 我们利用正确率、召回率以及 F1 测度来评价结果的好坏[9]:

$$\text{正确率(P)}=\frac{\text{系统正确发现的错误数}}{\text{系统发现的错误总数}}\times 100\%$$

$$\text{召回率(R)}=\frac{\text{系统正确发现的错误数}}{\text{文本中的错误总数}}\times 100\%$$

$$\text{F1 测度}=\frac{2\times \text{正确率}\times \text{召回率}}{P+R}\times 100\%$$

3 混淆词与介词搭配的方法

基于 Winnow 的算法使得系统整体的正确率、召回率以及 F1 测度分别达到 74.96%, 72.15%及 70.97%, 但是对于某些混淆词, 其正确率、召回率以及 F1 测度却很低. 例如: 混淆集(adapt, adept, adopt)在基于 Winnow 算法上经过测试, 得到 adapt(正确率: 66.67%, 召回率: 60%, F1 测度: 63.16%), adept(正确率: 55.56%, 召回率: 50%, F1 测度: 52.63%), adopt(正确率: 66.67%, 召回率: 80%, F1 测度: 72.72%).

针对这类正确率等指标较低的混淆词组, 我们可以通过设计其他的方法来进行改善. 通过仔细观察上面的这组混淆词我们可以发现, adapt 其后常跟介词“to”和“for”, adept 其后常跟介词“at”, 而 adopt 是及物动词后面直接接宾语无需加介词, 这样我们就可以利用混淆词与介词搭配的方法来检错:

第一步: 对于每个混淆词, 我们会从 BEC 网站抓取 40 条句子来进行训练, 10 条句子进行测试, 当一个句子中含有某个混淆词且混淆词后面紧接着一个介词, 那么就将这个介词写入这个混淆词所对应的介词向量中, 对每个混淆词都做这样的处理, 直到生成所有的介词向量. 当然, 如果含有某个混淆词的 40 个句子中, 混淆词后面都不接介词, 那么这个混淆词所对应的介词向量为空;

第二步: 如果在某组混淆词中(假设这组混淆词有三个混淆词), 某个介词出现在其中两个或者三个混淆词所对应的的介词向量中, 那么就把这个介词从其所出现的介词向量中删除, 对所有的混淆词组都做这样的处理;

第三步: 在对某组混淆词进行测试的时候, 如果一个句子的目标词后出现某个介词, 这个介词也在这组混淆词中某个混淆词所对应的的介词向量中出现, 那么就将这个介词对应的权重设的较大, 我们根据实验结果分析认为这个权重设为“3”较合适. 这时, 这个介词算作一个特征计入对应的混淆词的特征向量中, 接下来仍然按照 Winnow 算法中的$\sum_{i=1}^{n} w_i x_i$进行求和并判定.

以混淆集(adapt, adept, adopt)为例:

我们针对“adapt”, “adept”, “adopt”这三个词从 BEC 网站抓取 150 条句子, 每个词抓取 50 条, 我们从包含“adapt”的句子中提取的介词向量是(to, for, from), 从包含“adept”的句子中提取的介词向量是(at, in), 从包含“adopt”的句子中提取的介词向量是一个空向量, 那么我们就不需要从这三个介词向量中删除任何介词, 因为每个介词向量中的介词都不同于其他介词向量中的介词. 假设“adopt”对应的介词向量是(in, to), 那么就要把介词“in”从“adept”和“adopt”对应的介词向量中除去, 把介词“to”从“adapt”和“adopt”所对应的介词向量中除去, 其他的混淆集亦是如此. 在测试的时候, 如果目标词后出现介词“for”, 我们就把其所对应的权重设为“3”, 接下来还是按照 Winnow 算法中的权重求和公式进行判定. 根据实验结果我们得知: 如果出现介词“for”, “for”前的目标词应该是“adapt”, 这是由于“adept”和“adopt”所对应的介词向量中都不包含介词“for”, 并且介词“for”所对应的权重是“3”, 这是一个较大的数值, 在权重求和的过程中有决定性的作用.

加入混淆词和介词搭配的方法后, 混淆词组真词检错(adapt, adept, adopt)的各项指标达到以下水平: adapt(正确率: 72.72%, 召回率: 80%, F1 测度: 76.19%), adept(正确率: 75%, 召回率: 60%, F1 测度: 66.67%), adopt(正确率: 72.72%, 召回率: 80%, F1 测度: 76.19%).

4 英语真词错误检查系统的设计

4.1 系统整体设计

英语真词错误检查系统主要实现的是对英语作文中真词错误的检查. 由于这是一个研究性的课题, 所以我们采用了一些简化方法来实现和测试算法模型: 对英语作文中容易出现的真词错误单词, 我们选用 344 组混淆词, 对于每一个混淆词, 我们从英国国家语料库官方网站抓取 50 条句子来表示可能会出现该真词错误的文档,Winnow 算法的相关参数设置为: θ=1.1, β=0.8, α=1.3, 每个特征的初始权重为 0.2. 在此系统中, 主要实现以下功能模块:

(1) 语料获取模块: 对收集的 344 组混淆集创建 xml 文件进行存储, 然后针对所有的混淆词, 从 BEC 网站抓取 35000 多条句子并用 xml 文件进行存储.

(2) 语料预处理模块: 对每一个出现混淆词的句子, 进行分词和词性标注, 并以 XML 格式进行存储.

(3) 特征提取模块: 主要从语义方面提取特征, 语义方面使用文档频率和信息增益, 最终的特征单词用特征向量表示, 用文档频率提取的特征单词以 XML 格式进行存储.

(4) 权重训练与介词向量生成模块: 将句子语料分为两部分, 其中的 80%作为训练集, 剩余的 20%作为测试集. 使用 Winnow 算法对训练集中目标词所对应的特征向量中特征单词的权重进行训练, 并使用混淆词与介词搭配的方法生成介词向量, 将更新后的权重以 XML 格式进行存储, 将生成的介词向量存储在一个 txt 文件中.

(5) 真词检查模块: 使用 Winnow 算法及混淆词与介词搭配的方法对测试集中的句子进行测试, 并判断混淆词的出现是否正确. 这里没有将两种方法分开是因为混淆词与介词搭配的方法还是会用到权重, 最终还是要将该权重纳入 Winnow 算法的权重求和公式中去.

整个系统实现之后就可以运用到英语作文的检查中, 对于一篇作文, 我们需要对每句话都进行检查, 如果一句话中出现我们所提取的混淆词, 我们就需要检查这个词用的对不对, 若是正确的就不需要进行修改, 若是错的就需要替换成正确的单词, 如何进行检查并进行相应的替换就像前面算法所详细描述的.

4.2 系统结构流程图

图 1 描述了结合 Winnow 算法和混淆词与介词搭配的方法来实现英语真词错误检查的整过过程. 首先我们通过训练集进行特征的提取, 主要是提取目标词前后的单词, 并对每个混淆词生成一个介词向量, 接着我们利用文档频率及信息增益进行特征的筛选, 并进行权重的训练, 最后我们从测试集提取特征, 并结合 Winnow 算法及混淆词与介词搭配的方法进行测试, 实验结果的好坏利用正确率、召回率以及 F1 测度来评价.

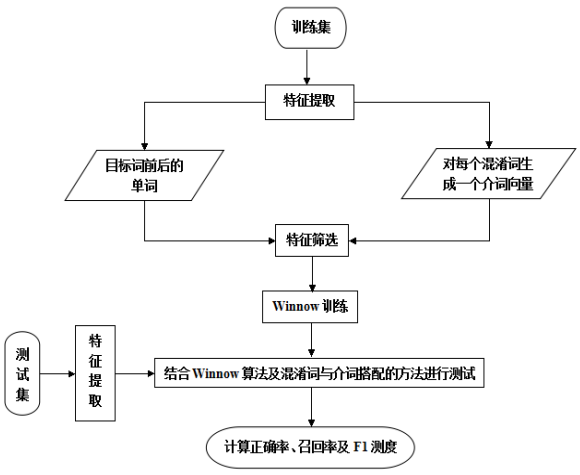

图 1　英语真词错误检查系统流程图

5　实验结果对比

表 1 是从加入混淆集与介词搭配的方法后各项指标都提高的混淆集中抽取的部分混淆词组. 我们从表中可以看出介词对一些混淆词组有较大的影响. 例如: 混淆词组(accede, exceed), “accede”是不及物动词, 后接宾语的时候要加介词“to”, 而“exceed”是及物动词, 后面直接接宾语; 因此当一个句子的目标词后出现介词“to”, 就可以判断目标词应该是“accede”而不是“exceed”. 这种混淆词与介词搭配的方法可以看做是一种规则的方法, 这与基于 Winnow 的统计算法结合, 可以有效地提高部分混淆集的正确率等指标.

表 1　实验结果

混淆集	基于 Winnow 算法			基于 Winnow 算法及混淆词与介词搭配的方法		
	正确率(%)	召回率(%)	F1 测度(%)	正确率(%)	召回率(%)	F1 测度(%)
accede	63.64	70	66.67	71.43	100	83.33
exceed	70	63.64	66.67	100	63.64	77.78
adapt	66.67	60	63.16	72.73	80	76.19
adept	55.56	50	52.63	75	60	66.67
adopt	66.67	80	72.73	72.73	80	76.19
allude	50	50	50	66.67	100	80
elude	70	70	70	100	70	82.35
allusion	53.33	80	64	60	90	72
delusion	71.43	50	58.82	85.71	60	70.59
herd	62.5	50	55.56	71.43	50	58.82
heard	58.33	70	63.64	61.54	80	69.57
hole	58.33	70	63.64	64.29	90	75
whole	66.67	54.55	60	85.71	54.55	66.67
imply	100	30	46.15	100	50	66.67
infer	66.67	60	63.16	85.71	60	70.59
precede	58.82	90.91	71.43	66.67	90.91	76.92
proceed	75	30	42.86	83.33	50	62.5
reluctant	58.33	70	63.64	75	90	81.82
reticent	62.5	50	55.56	87.5	70	77.78
singly	66.67	80	72.73	71.43	100	83.33
singularly	77.78	63.6%	70	100	63.64	77.78

6　结语

虽然基于 Winnow 的算法已经能够使系统整体的正确率、召回率及 F1 测度达到 70%以上, 但对于某些混淆集, 其最终测得的正确率等指标还是比较低, 为此我们引入混淆词与介词搭配的方法. 通过实验对比我们可以看到一些依赖于介词的混淆词在加入这种方法后各项指标有较大的提高, 系统的整体性能也更好. 接下来我们将进一步扩充语料库并增加混淆集的数量, 以便将系统应用到实际.

参考文献

1 Lawrence D. Spelling Checker and Corrector, 1992 .
2 Lee L. Similarity-Based approaches to natural language processing[Master’s Thesis]. Cambridge, MA: Harvard University, 1997: 35–56.
3 张仰森,俞士汶.文本自动校对技术研究综述.计算机应用研究, 2006, (6) : 8–12.
4 Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992, 24(2):377–439.
5 张磊,周明,黄昌宁,等.中文文本自动校对.语言文字应用, 2001,(1):19–26.
6 陆玉清,洪宇,陆军,姚建民,朱巧明.基于上下文的真词错误检查及校对方法.中文信息学报,2011,25(1):85–90.
7 Golding AR, Roth D. Apply winnow to context-sensitive spelling correction. Proc. of the 13th ICML. Bari, Italy. 1996.
8 Golding AR, Roth D. A winnow-based approach to context-sensitive spelling correction. Machine Learning, 1998, 34(1-3): 107–130.
9 李斌,姚建民,朱巧明.英文作文的自动拼写检查研究.郑州大学学报(理学版),2008,40(3):48–51.