文章编号: 1001 - 9081(2004)12Z - 0259 - 03

特征和语言模型结合的中文文本查错

陈笑蓉,秦 进
(贵州大学 计算机科学系,贵州 贵阳 550025)
(qin_gs@163.com)

摘 要:引入上下文词、搭配两种特征作为约束,利用最大熵原理构建查错模型,以期解决词语搭配不当、长距离的语言限制等词一级的错误。基本思想是,构建词语 y在上下文 x上的条件概率分布模型 $p(y|x)$,查错时根据句子中的上下文 x计算 y出现的条件概率 $p(y|x)$和 y不出现的条件概率 $p(\overline{y}|x)$。如果 $p(y|x) \geq p(\overline{y}|x)$,则 y无误;反之,y有错。实验结果表明,该方法获得了平均 91.14% 的召回率和 65.70% 的准确率。

关键词:查错;特征;语言模型;最大熵
中图分类号: TP391.12 **文献标识码**: A

0 引言

文本自动校对是指利用计算机自动识别文本中的拼写错误,并给出改正方案。所以,文本校对自然地有两个基本任务: 1)查错,即发现并警示出文本中的错误; 2) 纠错,即给出所警示错误的改正方案。一般来说,改正方案都不是唯一的,有多个改正方案供用户选择。

如果用 A表示文本中的错误总数,B表示系统警示的错误总数,C表示系统正确警示的错误数,D表示系统误判的错误数,则文本自动查错系统的主要评价指标如下:

召回率 $R = C/A * 100\%$,

准确率 $P = C/B * 100\%$,

误判率 $E = D/B * 100\%$。

现有中文自动校对的主要问题在于: 1)语言模型简单。大多数自动校对方法使用字、词或词类的 N元模型。但是 N元模型只反映局部的语言限制,不能包括长距离的语言限制。2) 多数基于词性 N元模型的校对方法都是先对文本进行自动词性标记,以解决一个词具有多个词性标记的问题,然后再利用自动标记相邻词性间的关系来检查文本中是否有错误。然而,根据这种自动标记的词性来进行检查的方法会掩盖许多可能的错误,甚至陷入"先有鸡还是先有蛋"的怪圈。3)目前大多数自动校对方法都是只使用字、词或词类的 N元模型中的一种模型,或者是孤立地分别使用其中的几种模型。而真正综合使用多种模型时,必须应用模型合一的过程或方法[1]。

针对上述问题,本文引入上下文词、搭配两种特征,利用最大熵原理构建查错模型,以期解决词语搭配不当、长距离的语言限制等词一级的错误。

1 最大熵语言模型

1950年 E. T. Jaynes首次提出最大熵方法,1992年 Della Pietra等人首先把最大熵原则应用于语言建模。最大熵语言模型被用于词性标注、文本切分、介词短语附着问题、句子边界检测、部分句法分析以及语音识别的语言建模等。

直观上最大熵原则是简单的:对知道的一切建立模型,对不知道的不作任何假设。换句话说,给定一个事实的集合,选择的模型的概率分布和所有事实一致,除此之外,其他方面模型的概率分布尽可能地平均。

一个随机过程的所有输出值构成有限集 Y,每个输出值 $y \in Y$。随机过程生成每个 y时,都要受到上下文信息 x的影响,所有与 y有关的上下文信息 x构成集合 X。最大熵模型用来表示随机过程的行为,它用来估计给定上下文 x的情况下,随机过程输出 y的条件概率。随机过程的模型 $p(x,y) \in P$。

构建的随机过程的模型 $p(x,y)$要和观察到的语言知识一致,这里,语言知识用特征的形式来表示。特征定义为一个二值函数 $f: X \times Y \rightarrow \{0,1\}$。特征 f关于经验概率分布 $\tilde{p}(x,y)$的期望值记作 $E_{\tilde{p}} f$,关于模型 $p(x,y)$的期望值记作 $E_p f$。建立的模型应该与从训练样本中观察到的事件一致,即要求:

$$E_p f = E_{\tilde{p}} f \tag{1}$$

等式(1)叫作约束等式,或者简称约束。

最大熵模型就是:

$$p^* = \arg\max_{p \in C} H(p) \tag{2}$$

其中,$C \equiv \{p \in P \mid E_p f_i = E_{\tilde{p}} f_i\}$, $\forall i \in \{1, 2, \cdots, n\}$, p^*的求取已经有成熟的算法,这里就不赘述。

2 中文文本查错

2.1 两种特征

特征是从语料中抽取的语言知识,特征向量可以用来表征文本。这里,特征作为文本证据被用于校对,即特征或者作为某种语言现象出现的证据,或者作为某种语言现象不出现的证据。两类特征证明对校对是有用的[3]:第一类是易错的目标词语左右 $\pm k$个上下文词;第二类是易错的目标词附近 l个连续的词和词性模式——搭配。

上下文词和搭配两种特征互为补充:前者是目标词所处的语言环境,记录了部分语义信息(篇章主题、时态等);后者则记录了局部的语法信息。$\pm k$叫作上下文窗口范围,k的取

收稿日期: 2004 - 02 - 11;**修订日期**: 2004 - 05 - 09 **基金项目**:贵州省科学技术基金项目(993021)

作者简介:陈笑蓉(1954 -),女,贵州贵阳人,副教授,主要研究方向:应用系统软件、中文信息处理、语义 Web; 秦进(1978 -),男,贵州人,助教,主要研究方向:应用系统软件、中文信息处理、文本挖掘.

值应该适中。窗口开小了,解决问题需要的信息量会不足,结果不佳;窗口开大了,会增加计算量,而且可能会引入噪声,也会影响结果。文献[4]提出"合适范围"应该满足下面的定性标准:在能够解决问题的前提下,提供的信息量足够大,产生的噪声足够小,并且有利于提高时空效率。作者通过实验得出结论,汉语词语的上下文窗口的合适范围是$[-8, TIF, +9]$,这个范围包含了85%的信息量。据此,本文让$k=8$,l的取值和文献[3]一样,即$l=2$。

2.2 特征选择

构建统计模型需要进行特征选择,这是因为存在于实际训练数据中的特征空间可能非常大,但并不是所有的特征信息都对随机过程的输出有影响。因此,通过特征选择压缩特征空间,可以减少计算量和降低噪声。

本文采用的特征选择策略是:对于上下文词特征,首先利用启发式规则选择出现次数大于10的上下文词,然后在这些上下文词中选择和目标词互信息大于阈值τ的特征。上下文词w与目标词t的互信息定义如下:

$$MI(w,t)=\log\left(\frac{p(w,t)}{p(w)\times p(t)}\right) \tag{5}$$

其中$p(w)$,$p(t)$分别是词w和t在语料中出现的概率,$p(w,t)$是w和t在语料中同时出现的概率。设语料的容量为N个词,$freq(x)$为x在语料中出现的次数,概率用最大似然估计MLE(Maximum Likelihood Estimate)方法来计算,则:

$$\begin{aligned} p(w)&=freq(w)/N\\ p(t)&=freq(t)/N\\ p(w,t)&=freq(w,t)/N \end{aligned} \tag{6}$$

互信息体现了两个词语的相关程度:互信息越大,词语的相关程度越高;反之,词语的相关程度越低。对于搭配特征,仅仅简单地选择出现次数超过10次的搭配模式。

2.3 查错模型

词语y在上下文x下的最大熵条件概率模型$p(y|x)$就是我们的查错模型,我们将利用最大熵原理求取$p(y|x)$。设t是句子中待判定的词语,于是查错函数$Check(t)$可以定义为:

$$Check(t)=\begin{cases} true & p(t|x)\geq p(\neg t|x)\\ false & \text{否则}\end{cases} \tag{7}$$

其中$p(\neg t|x)$是在x下输出不为t的概率。$Check(t)$为$true$说明t在句子中的出现是合法的;反之,t的出现是不合法的。

一个汉语句子$S=c_1c_2\cdots c_n$,其中$c_i\in GB$,GB是国标汉字字符集,包括6763个汉字、英文字母、阿拉伯数字以及标点符号等。经过切分后,$S=w_1w_2\cdots w_m$ $(m\leq n)$,其中$w_i\in D$,D是词典。目标词t的一个上下文x表示为词语组成的向量,即$x=(w_{-k},\cdots,w_{-2},w_{-1},w_1,w_2,\cdots,w_k)$。模型的输出$y\in Y=\{t,\neg t\}$,其中$\neg t$表示输出的词语不是$t$。$(x,y)$称为一个事件,$X\times Y$称为事件空间。

模型的一个上下文词特征是一个二值函数$f: X\times Y\rightarrow\{0,1\}$。

$$f_{c,t}(x,y)=\begin{cases}1 & c\in\{w_{-k},\cdots,w_{-2},w_{-1},w_1,w_2,\cdots,w_k\},\text{且 } y=t\\ 0 & \text{否则}\end{cases} \tag{8}$$

其中$x\in X$,$y\in Y$。

模型的一个搭配特征是一个二值函数$f: X\times Y\rightarrow\{0,1\}$。

$$f_{c,t}(x,y)=\begin{cases}1 & \exists i(i=\pm1,\pm2),\text{使得 } c=Tag(w_i),\text{且 } y=t\\ 0 & \text{否则}\end{cases} \tag{9}$$

其中$x\in X$,$y\in Y$,$Tag(w_i)$是w_i的词性标记。

3 实验及结论

3.1 实验数据

实验从训练语料中选取19个出现次数较高的词语作为测试数据,表1第2列和第3列分别给出了这些词语和相应的出现次数。实验测试这些词语在句子中出现时是否合法。

词表 词表用于查错的预处理过程——切分。实验使用的词表规模有60 450个词条,每个词条包括词和词性两个属性。

训练语料 1998年1月一个月的已经经过切分、标注的《人民日报》熟语料,用于抽取特征和训练特征的权值。从中仅仅选取每个测试词语出现的句子作为训练样本。

测试语料 2000年《人民日报》生语料,从中选择需要测试的词语的测试句子。

测试句子 由于真实的错误句子很难收集,实验测试采用人工构造的句子。为每个词语构造测试句子的方法是:从测试语料中选取出现该词语的句子若干作为测试句子的一部分,再从测试语料中选取相同数量的和该词语易混淆的词语的句子,把这部分句子中的相应词语替换成该词语后得到的句子作为测试句子的另一部分。这里的易混淆是指同音易混淆。从测试语料中选取的每个词语的测试句子数如表2第3列所示。

表1 测试数据

编号	查错词语	语料中出现的次数	训练样本数	特征数
0	经济	2 657	2 137	337
1	记者	2 143	1 999	233
2	一个	1 889	1 704	35
3	北京	1 377	877	150
4	改革	1 282	876	655
5	政府	1 225	1 206	635
6	公司	1 157	689	193
7	世界	1 154	841	220
8	领导	1 134	804	477
9	关系	1 049	714	281
10	国际	1 008	685	164
11	技术	938	604	250
12	自己	932	740	422
13	职工	893	488	104
14	生产	883	610	267
15	同志	878	615	322
16	加强	844	685	394
17	会议	834	633	171
18	组织	830	679	223

3.2 实例分析及实验结果

3.2.1 实例分析

假设待查错的句子为:

单说这草垛山有大片原始森林,林中经济丛生,有黄羊、野猪、虎、豹、豺、狼等出没。

经过切分后,抽取得到该句子的上下文(窗口为±8)特

征向量 x = (草垛,山,有,大片,原始,森林,林,中,丛生,有,黄羊,野猪,虎,豹,豺,狼)。其中没有一个支持“经济”出现的特征,全是支持“¬经济”出现的特征。据此,计算得到 $p(经济 \mid x) = 0.076103 < p(\neg 经济 \mid x) = 0.923897$,所以“经济”在该句子中的出现是不合法的。

3.2.2 实验结果

训练阶段得到的每个测试词语的特征数如表 1的第 5列所示。19个词语的测试结果如表 2所示。大部分词语的召回率在 80%以上,准确率在 50%以上。平均召回率 91.14%,平均准确率 65.70%。

从表 1和表 2可以看出规律,召回率低的词语,特征集大;准确率低的词语,特征集小。可见,查错效果的好坏和特征集的大小密切相关。在召回率和准确率之间一个好的折中,需要适合的特征集大小。以“一个”为例,其在不同的特征集大小下的查错效果如图 1所示。可以看出随着特征集增大,召回率急剧下降,从 50个特征时的 97.19%下降到 500个特征时的 24.29%,下降了 73个百分点;而准确率随着特征集的增大,增长很缓慢,从 50个特征时的 51.93%增长到 500个特征时的 60.81%,只增长了 9个百分点。

表 2 测试结果

编号	测试词语	测试句子数	警示错误的句子数	正确警示的句子数	召回率 (%)	准确率 (%)
0	经济	(796 +25) * 2	1159	785	95.61	67.73
1	记者	(23 +48 +76) * 2	163	137	93.19	84.04
2	一个	(101 +327) * 2	832	422	98.59	50.72
3	北京	757 * 2	1309	749	98.94	57.21
4	改革	(25 +29) * 2	47	44	81.48	93.61
5	政府	(106 +421) * 2	579	421	79.88	72.71
6	公司	(33 +108 +43 +41 +1060) * 2	2270	1266	98.52	55.77
7	世界	(220 +194) * 2	607	346	83.57	57.00
8	领导	142 * 2	158	110	77.46	69.62
9	关系	912 * 2	1116	845	92.65	75.71
10	国际	(65 +15) * 2	137	78	97.50	56.93
11	技术	(70 +59 +38 +23 +26) * 2	336	207	95.83	61.60
12	自己	(89 +33 +31) * 2	179	119	77.77	66.48
13	职工	(177 +35) * 2	403	212	100	52.60
14	生产	(22 +59) * 2	134	78	96.29	58.20
15	同志	(1584 +328) * 2	2514	1616	84.51	64.28
16	加强	113 * 2	135	113	100	83.70
17	会议	335 * 2	507	323	96.41	63.70
18	组织	(47 +32 +229) * 2	454	257	83.44	56.60

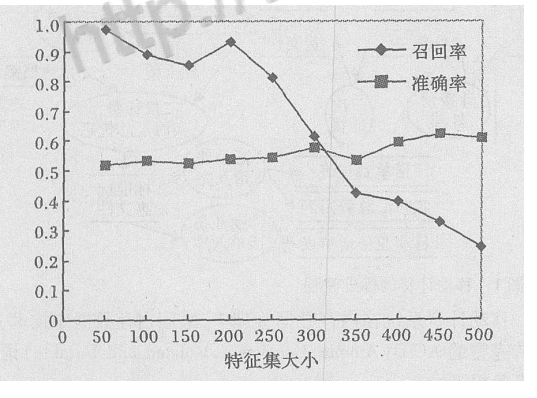

图 1 特征集大小和查错效果的关系

4 结语

实验证明本文提出的方法是有效可行的,与已有的其他方法比较存在几个优点:

1) 采用的上下文词和搭配两种特征覆盖了长距离的语言限制和局部的语言现象,把原有的利用字、词的接续信息查错提升到利用字、词的语义、语法信息查错,使得查错的深度更进了一层。

2) 最大熵语言模型与其他语言模型相比,它能把各种异构的语言知识以特征的形式加以统一应用,并对未知的知识保留了最大的不确定性。也就是说,最大熵方法同时解决了知识表示的问题,可以很方便地随时把新获取的语言知识添加到模型中去,不断提高查错的效果。

3) 部分解决了数据稀疏的问题,使得在训练语料规模较小的情况下也能取得较好的查错效果。

参考文献:

[1] 张磊,周明,黄昌宁,等. 中文文本自动校对[J]. 语言文字应用, 2001, (1).

[2] BERGER AL, STEPHEN A, DELLA PIETRA SA, *et al* A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39 - 71.

[3] GOLDING AR. A Bayesian hybrid method for context - sensitive spelling correction[A]. Proceedings of the Third Workshop on Very Large Corpora[C]. Cambridge, MA, 1995. 39 - 53.

[4] 鲁松,白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报, 2001, (7): 742 - 747.

[5] LAU R, ROSENFELD R, ROUKOS S. Adaptive Language Modeling Using the Maximum Entropy Principle[A]. Proceedings ARPA Human Language Technologies Workshop[C]. Princeton, NJ, 1993. 81 - 86.

[6] 陈笑蓉,秦进,汪维家,等. 中文文本校对技术的研究与实现[J]. 计算机科学, 2003, (11).

阅读此文的还阅读了：

1. 赏析文本语言 培养语言能力
2. 直面文本 利用文本 发展语言
3. 特征和语言模型结合的中文文本查错
4. 如何结合文本开展语言知识的教学
5. 《狗心》文本语言浅析
6. 基于语言模型的神经网络的文本情感分析
7. 基于语料库的中文自动查错综述
8. 亲近语言走进文本
9. 基于语言模型的藏文文本分类研究
10. 语言走进文本
11. 中文自然语言处理与计算机的结合问题研究
12. “品味语言”为视角的初中文言文文本解读策略探索
13. 紧扣文本语言特质,实施读写高效结合
14. 减负 从选中文本后开始
15. 文本挖掘与中文文本挖掘模型研究
16. 走进文本,品析语言
17. 基于规则与统计相结合的中文文本自动查错模型与算法
18. 文本·语言·意蕴
19. 语篇教学中文本语言的处理策略
20. 基于知识库的多层级中文文本查错推理模型
21. 中文“词”的语言模型识别研究方法综述
22. 英语课本中真实语言模型和人工语言模型的结合
23. R语言中文分词方法在审计短文本中的应用
24. 《意思文本》模型的深层句法语言
25. 基于统计语言模型的中文自动文本分类系统

26. 赏析文本语言 培养语言能力

27. 中文文本自动校对的语义级查错研究

28. 基于查询扩充机制的中文文本检索模型

29. 把握语文教材中文本语言“真实表意”的方法

30. 挖掘文本语言 拓展语言训练

31. 初中文本阅读的迁移路径

32. 紧扣文本语言特质提高读写结合效率

33. 你会查错吗

34. 基于类比语料库的中英旅游文本对比研究——语言特征和文本功能维度

35. 走进文本,品析语言

36. 幼儿园语言教学中文本价值挖掘的研讨

37. 语文课程中文本语言与学生经验之矛盾探析

38. 高中文本阅读的语言分析策略

39. 立足文本,感悟文本语言

40. 让语言品味回归文本

41. 分析Python语言的中文文本处理

42. 涵泳文本语言挖掘文本内涵

43. 基于语言模型的中文文本分类系统

44. 语言走进文本

45. 一种从自然语言文本到本体模型的转换方法

46. 贝叶斯同语言模型相结合的中文文本分类方法的研究

47. 品读,让文本语言更有“度”

48. 论新闻写作中文本语言的仪式化表述

49. 中文分词系统及中文文本的分词方法

50. 基于二元、三元统计模型与规则相结合的中文文本自动查错研究