Approach in automatic detection and correction of errors in Chinese text based on feature and learning

Zhang Lei[1] Zhou Ming[2] Huang Changning[2] Lu Mingyu[1]

1:Department of Computer Science & Technology, Tsinghua University, Beijing, 100084

2:Microsoft Research China, Beijing, 100080

Abstract: Language models adopted by most existing error detection and correction approaches of Chinese text are N-Gram models of character, word or POS tag. Their deficiencies are that only local language constraint is employed and there is no language model unification process. A feature-based automatic error detection and correction approach is presented. It uses both local language features and wide-scope semantic features. Winnow is adopted in the learning step. In experiment, this method achieves error detection recall rate of 85%, precise rate of 41%, and error correction rate of 51%. It shows better performance than existing approaches based on N-Gram models.

Keywords: Automatic error detection and correction of Chinese text; Natural language processing; Spelling check

基于特征与学习的中文文本自动校对方法*

张磊[1] 周明[2] 黄昌宁[2] 鲁明羽[1]

1:清华大学计算机科学与技术系，北京，100084

2:微软中国研究院，北京，100080

摘要：现有的中文自动校对方法使用字、词或词类的 n 元语言模型。它们的问题在于：仅使用某种局部语言现象的统计特征，导致处理能力不足；多种语言模型没有合一。本文设计实现了一个基于特征的中文自动校对方法。该方法综合考虑了中文文本中字、词和词性的局部语言特征以及长距离的语义特征，并且采用 Winnow 方法进行特征学习。评估实验表明该方法召回率达到 85％，准确率达到 41％，订正率达到 51%。性能比目前常用的词的 N 元模型方法有明显的提高。

关键词：中文自动校对；自然语言处理；拼写检查

现有的中文自动校对方法[2,3]的召回率一般小于 70%，准确率小于 40%。它们的主要问题在于：⑴语言模型简单。大多数方法使用字、词或词类 n 元语言模型。但是 n 元模型只反映局部的语言限制，而不包括长距离的语言限制。⑵多数基于词性 n 元模型的校对方法都先对文本进行自动词性标注来解决一个词具有多个词性标记的问题，然后再利用自动标注后相邻词性间的接续强度的大小来判断文本中是否有错。然而现有的自动词性标注过程是试图找到一条具有最大词性标记接续强度路径的过程，因此这样的校对方法会掩盖许多可能的错误。另外，使用与标注过程相似的判断依据去判断标注结果的合理性也陷入了“先有鸡还是先有蛋”的怪圈。⑶多数方法都是只使用字、词或词类的 n 元模型中的一种模型，或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，应该有模型合一的过程或方法。

针对这些问题，我们设计了一个基于特征[4,5,7]的中文自动校对方法。所谓特征就是在特定目标（字或词）的上下文中出现的语言现象。基于特征的方法有两个优势：首先，特征可以各式各样，采用什么样的模板来提取特征可以根据不同的应用确定。通过不同的特征模板，既可以利用到局部的语言限制，也可以引入长距离的语言限制。其次，将不同模板的特征统一对待，可以实现语言模型合一的功能。本方法综合使用了词的 2 元接续关系、词性类的 3 元接续关系、上下文语义类、词内邻接汉字等特征模板。在特征学习上，由于特征空间的规模非常巨大，并且特定的目标只和整个空间中的极少数特征相关，所以我们采用 Winnow[6]方法对特征和目标之间联系的权值进行学习。经过 200M 语料训练后，试验显示其召回率达到 85％，准确率达到 41％，订

* 本课题为国家教委博士点基金项目

正率达到 51%。

本文详细介绍了特征学习模型和我们的自动校对方法。第 1 节描述中文校对的模型和特征模板；第 2 节仔细讲述特征学习的 Winnow 模型，包括学习过程和校对过程；第 3 节给出实验结果；最后是结论。

1、校对任务描述

本文使用“字串”特指单个汉字或词。字串 w 的混淆集记作 $cfs(w)=\{y_1,y_2,...,y_k\}$，其中 $y_i \neq w$。它表示错误文本中的字串 w 在正确的原文中可能是以字串 y_i 的形式出现。也就是说，w 可能是由 y_i 发生歧义造成的错误。利用混淆集，我们将校对问题转化成排歧问题。设一个经过分词的句子 $S=w_1w_2...w_n$，其中 w_i 是切分后的词。试图判定 w_i 是否正确时，校对的任务就是从 $\{w_i\}\bigcup cfs(w_i)$ 中选择一个最适合上下文的字串 $\tilde{y}$。如果 $\tilde{y}\in cfs(w_i)$，则表示算法认为 w_i 出现在此处上下文中是个错误，并且应该被修改成 $\tilde{y}$。如何构造混淆集本身是一个有趣的课题。实验中我们仅根据五笔字型输入编码相近的原则来构造混淆集。例如，汉字“温”(五笔字型输入码为 ijlg)的混淆集是{漫 (ijlc)，湿(ijog)，汽车(irlg)...}。

排歧的过程是一个评价选择的过程。需要评价的是句子中上下文对特定目标的支持度。在基于特征的方法中，句子中指定目标的上下文用一组特征的集合来表示，称为活跃特征集。它由专门的特征提取器根据指定的特征模板从目标的上下文中提取。设目标字串 x_i 在分词后落在词 W_j 中，其中 x_i 可以是两字或多字词 W_j 中的一个汉字，也可以是 W_j 本身。针对中文的特点，我们使用下面几种特征模板来提取特征：

- 词的 2 元接续关系：这个模板记录目标字串所在词的前后词。这类特征形如 $W_{j-1}(*)$ 和 $(*)W_{j+1}$。其中 $(*)$ 表示当前目标字串所在的词。
- 词性类的 3 元接续关系：如引言所述，目前校对系统使用词性信息时应该绕过自动词性标注环节。为此，我们引入词性类的概念。设所有可能的词性标记组成的集合为 T，则 T 的所有非空子集都是一个词性类，并被赋予唯一的词性类标记。简单地说，具有相同的可能词性的词属于同一词性类。例如有且仅有名词和动词两种可能词性的词的词性类标记都是动名词性类。这类特征形如 $C_{j-2}C_{j-1}(*)$、$C_{j-1}(*)C_{j+1}$ 和 $(*)C_{j+1}C_{j+2}$。其中 $(*)$ 表示目标字串所在词 W_j 的词性类标记。C_i 表示词 W_i 的词性类标记。
- 上下文语义类：本文将目标字串所在词 W_j 上下文中的 $W_{j-k},W_{j-k+1},...,W_{j-2}$ 和 $W_{j+2},W_{j+3},...,W_{j+k}$ 等词的语义类标记作为一种特征。其中 $k\geq 2$ 为常数，表示上下文窗口的范围。本文中 k 取 6。由于前面已经将 W_{j-1}、W_{j+1} 作为特征，所以这两个词的语义类就不再作为特征出现了。我们基本上按照《同义词词林》[1]中的第 5 层的语义类标记作为各个词的语义类标记。对于具有多个语义类标记的词，我们采用与词性类相似的办法，根据它们的具有的第 1 层标记的组合赋予它们一个新的语义类标记。
- 词内邻接汉字：一个错误的两字或多字词，它既可能是由其它汉字或词一下错成的，例如“*他/喝/了/一/碗/粥*”错成“*他/喝/了/一/碗/继续*”；也可能是由词内某个汉字的错误造成，如“*他/要/去/上海*”错成“*他/要求/上海*”。对后一种错误，“要求”的混淆集中不会有“要 去”这样的字串。如果想正确地发现这种错误，就必须另外考虑“要求”中每个汉字出错的情况，即考虑“求”是“去”出错造成的情况。前三种模板以“求”为目标提取上下文特征时，实际上是以它在切分后所在词“要求”为目标提取的。为了反映“求”和“要求”的差别，对单个汉字，我们还提取它的词内邻接汉字特征。设目标汉字 x_i 分词后所在词 $W_j=x_{i-p}...x_i...x_{i+q}$，则这种特征形如 $x_{i-1}(*)x_{i+1}$。其中 x_{i-1}、x_{i+1} 分别是 x_i 在词 W_j 内的前后邻接汉字。如果 x_i 是 W_j 的词头，则 x_{i-1} 为 ε；如果 x_i 是 W_j 的词尾，则 x_{i+1} 为 ε。

词的 2 元接续关系特征反映了目标字串附近词的接续关系。词性类的 3 元接续关系特征反映了目标字串附近的局部句法搭配。上下文语义类特征则能包含目标字串两侧较长范围里的语义类信息。这三种特征互补性很强。词内邻接汉字特征则是针对中文文本中词与词之间没有明显的分隔符号的特点而专门设计的。

2、算法设计

图 1 简单表示了区分“温”和 “漫”的算法模型。中间部分标为“漫”和“温”的椭圆表示这两个目标的分类器 $\theta_x(F)$。对给定的目标字串 x 和从上

下文中提取出的活跃特征集 F，$\theta_x(F)$ 取值为 0 表示该分类器认为字串 x 不适合所给出的上下文。否则 $\theta_x(F)$ 取值为 1。从图中可以看到，每个分类器都与底部的许多上下文特征相连。每个连接都有一个权值 w。具体地说，对目标 x 和上下文中提取出的活跃特征集 F：

$$\theta_x(F)=0 \quad \Leftrightarrow \quad \sum_{f\in F} w(f,\theta_x) < \varepsilon$$

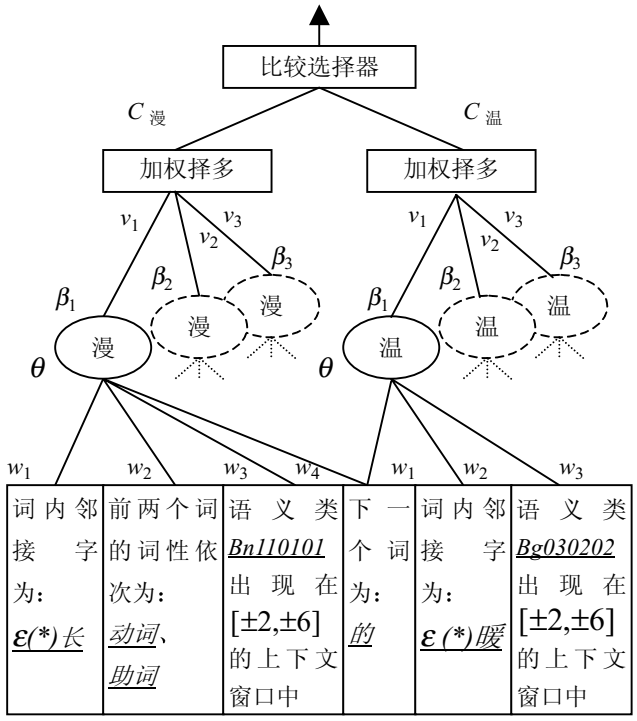

图 1. 区分汉字“温”和“漫”的模型示例

其中 $\varepsilon>0$ 为常数阈值，本文取 1.0。$w(f,\theta_x)$ 表示特征 f 和分类器 θ_x 之间连接的权值。通常，特征空间的规模非常大，而具体的一个分类器只和其中极少数的特征相连。那些不存在的连接的权值被认为是 0。而存在的连接的权值都是经过学习获得。图中用虚线椭圆框表示的分类器是使用不同的学习参数习得的。在学习和校对过程中，使用不同参数的分类器会有不同的表现，让表现好的分类器得到较多的信任是合理的。所以目标 x 的每个分类器 θ_x^j 也都有相应的权值 v_j 表示该分类器的可信程度。θ_x^j 表现越好，v_j 也越大。图中标有“加权择多”的方框部分对目标 x 的各个分类器的结果进行加权并得到的最终评价结果 $C_x(F)$，它反映了活跃特征集 F 对目标 x 的支持程度。

2.1 学习过程 学习过程包括：建立特征和目标之间的连接；更新连接的权值；分类器的可信度设定。

2.1.1 建立连接 第一步需要学习分类器与上下文特征之间的连接。首先对训练语料中的句子分词。然后，对分词结果中的每个词 W_k，分别以 W_k 和 W_k 中每个汉字为目标从分词后的句子中提取出活跃特征集 F，并为目标的每个分类器与 F 中的每个特征建立连接，并给这些连接赋初始权值。本文所有的初始权值都取 0.1。

2.1.2 权值更新 对一个切分后的句子中的的每个词 W_k：⑴我们称以 W_k 为目标提取出的上下文活跃特征集 F 是 W_k 的一个正例，同时又是 $\forall y\in cfs(W_k)$ 的反例。⑵如果 W_k 是个两字或多字词，则依次考察其中的每个汉字 x_i，称以 x_i 为目标提取出的上下文活跃特征集 F 是 x_i 的一个正例，同时又是 $\forall y\in cfs(x_i)$ 的反例。分类器对正例的判定结果应该为 1，对反例的判定结果应该为 0。如果分类器作出了与此相反的判断，就需要调整特征和分类器之间连接的权值：

- 当分类器 θ_y 对正例 F 取值为 0 时，对 $\forall f\in F$，需要提高它与该分类器之间连接的权值，即令 $w(f,\theta_y)=\alpha\cdot w(f,\theta_y)$。其中常数 $\alpha>1$，称为权值提升参数。
- 当分类器 θ_z 对反例 F 取值为 1 时，对 $\forall f\in F$，需要降低它与该分类器之间连接的权值，即令 $w(f,\theta_z)=\beta\cdot w(f,\theta_z)$。其中常数 $0<\beta<1$，称为权值降低参数。

权值更新过程同样要扫描一遍语料。对其中每个句子切分后的每个词 W_k，先用 W_k 的分类器对作为正例的活跃特征集 F 进行判断，如果判断结果为 0，则提升有关连接的权植；然后用 $cfs(W_k)$ 中每个元素的判定器对作为反例的该活跃特征集 F 进行判断，如果判断结果为 1，按上述方法降低权值。当 W_k 是两字或多字词时，还要分别用 W_k 中的每个汉字 x_i 以及其混淆集 $cfs(x_i)$ 中的各个元素的分类器判定提取出的活跃特征集，并在判定出错时调整权值。

2.1.3 分类器可信度设定 参数 α、β 的选择对分类器的性能有较大的影响。通过给参数 α、β 设定不同的值，可以为指定目标 y 构造多个分类器 θ_y^J。实验中我们设定 $\alpha=1.5$，用不同的 β 值 0.90、0.75 和 0.65 构成不同的分类器。对不同的分类器 θ_y^J，根据它们在学习中表现出的性能给予不同的信任 $v_j=\gamma^{m_j}$。其中 $0<\gamma<1$ 是常数。m_j 是分类器 θ_y^J 在学习中做出错误判断的总次数。在本文中

$\gamma = 0.98$。加权择多过程对不同参数分类器的判定结果进行综合评分，结果为：

$$C_y(F) = (\sum_j v_j \theta_y^j(F)) / (\sum_j v_j)$$

2.2 校对过程 对切分后的待查错句子 $S = W_1 W_2 ... W_n$ 中的每个词 W_k，首先分别用 $\{W_k\} \bigcup cfs(W_k)$ 中的每个字串 y 替换 W_k，得到句子 S_y^k。对新句子重新分词并保证 y 不被切碎。然后以新句子中的 y 为目标提取上下文活跃特征集 F_y^k，根据 Winnow 模型计算出 $C_y(F_y^k)$。最后，选择出最适合上下文的字串 $\tilde{y} \in \{W_k\} \bigcup cfs(W_k)$，使得

$$\sim(F_{\sim}^{k}) = \underset{\in\{\ _k\}\cup\ \{\ _k)}{MAX} \qquad (F^{k})$$

在下面两种情况下，原待查错句子中的 W_k 被认为有错：

- $\tilde{y} \neq W_k$。

  此时，我们还认为 W_k 应该被改正成 $\tilde{y}$。

- $\tilde{y} = W_k$，但 $_k\ F_{\ k}^{k}) < \psi$。

  其中常数 $0 < \psi < \varepsilon$。本文取 $\psi = 0.3$。此时虽然 W_k 比其混淆集中所有的字串都更适合句中的上下文，但它所获得的绝对支持度太低，所以我们也认为它有错。只不过我们无法给出改正方案。

如果 W_k 为两字词或多字词且上面步骤没有判断 W_k 有错，则我们还要用同样的方法判断 W_k 中的每个汉字是否有错。

3、实验结果与分析

实验使用的词表是《同义词词林》中的词表与一部 80000 带词性标记的词表的交集。按照第 1 节描述的方法，其中的每个词都标记了语义类和词性类。训练语料为 93、94 年的人民日报、94 年的市场报和 94 年的百家报，总共约 1 亿字。首先先定义自动校对的几个评价指标，令：

A = 文本中的错误总数

B = 校对方法警示的总数

C = 校对方法正确侦测出的错误数

D = 校对方法正确侦测出并正确订正的错误数

则：召回率=*C/A**100%，准确率=*C/B**100%，订正率=*D/A**100%。

考虑到获得真实的错误文本的困难性以及 Winnow 方法学习过程所需要的时间和空间开销，我们的测试分为伪错误测试和真实错误测试两部分。

错误类型	错误数	标记出的错误数	查错召回率(%)
混淆集替换错	1119	1070	95.6
随机替换错	191	170	89.0
缺字错	125	21	16.8
重字错	60	12	20.0
加字错	60	50	83.3

表 1 各种伪错误检查结果

在伪错误测试中，我们随机选择了 5 个字串作为检查目标并生成了它们的混淆集。这 5 个目标是：“温”，“罕”，“做”，“模式”和“信念”。对每个目标，我们都对从语料库中随机选择的含有该目标的正确句子和利用语料库中的句子随机生成的错误例子用我们的方法进行判断。随机生成的错误有以下几种形式：⑴对含有目标混淆集里元素的句子将该混淆元素用检查目标替换。称这种错误为混淆集替换错。⑵将任意句子中随机的一个汉字用检查目标替换。称为随机替换错误。⑶在含有检查目标的句子中删除目标左边或右边的汉字。称为缺字错误。⑷对含有单字检查目标的句子将该目标汉字重复一次。称为重字错误。⑸在任意句子的随机位置上增加一个单字检查目标。称为加字错误。为保证测试文本中错误率接近真实的错误文本，随机选择的正确句子的总数控制在错误句子的 20 倍左右。实验共检查了 31180 个正确的句子，其中有 1905 个句子被标记为有错。本方法对各种错误句子判断的结果见表 1。除此之外本方法对正确标记出的 1070 个混淆集替换错中的 796 处给出了正确的纠正方案。试验表明其召回率达到 85%，准确率为 41%，订正率为 51%。作为比较，当使用传统的词的 3 元模型对上面的伪错误进行测试时，召回率只有 67%，准确率为 28%。

在真实错误的测试中，我们对含有 335 个真实错误的以五笔字型输入法录入的文本进行了测试。本校对方法总共标记了 453 处可能有错的地方，其中正确标记出的错误有 291 处，召回率达到 88%，准确率达到 64%。在检查出的错误中，有 189 处还给出了正确的改正方案，订正率为 56%。作为比较，当使用词的 3 元模型校测试时，查错的召回率和准确率分别只有 69%和 35%。

可见，Winnow 方法比原有的单纯使用 n 元模型的方法在查错的召回率和准确率方面都有比较明显

的提高，尤其是召回率。这是因为：⑴上下文语义类特征的引入使一些无论使用字、词还是词性的 n 元模型都很难判断出来的有错句子能够被检查出来。例如下面的有错句子(下划线标记出错误部分，方括号内是其正确原文)就被我们的方法正确地判断出来：

进一步研究观察发现，当空气中的[湿]温度稍微大一些时，蜘蛛...

我虽然没有[介入]做过他们的争执，...

在我们国家，广播仍然是极其重要的传播[形式]模式，...

⑵综合使用多种互补性强的上下文特征弥补了单一特征或模型查错能力的不足。⑶混淆集的概念将校对过程近似为排歧过程，这使查错更加有的放失。⑷学习效果良好的 Winnow 模型保证了排歧过程过程的准确性。

数据稀疏是本方法遇到的一个问题。当混淆集中存在出现频度非常高的字串时，那些由混淆集中出现频度比较低的元素产生的替换错误在实验中常常被错误地建议成高频字串。例如一些由“敌”、“介入”、“优秀”等字串替换成“做”的错误被建议改成高频度的“作”。另外，算法的空间开销非常大，因为我们存储的是目标与特征之间的一个稀疏矩阵。

4、结论

本方法实现了基于特征和 Winnow 学习模型的中文自动校对方法。它有以下特点和优势：⑴综合使用词的 2 元接续关系、词性类的 3 元接续关系、上下文语义类、词内邻接字等四种特征模板。它们的互补性很强；既包含了局部的语言限制，也包括了长距离的语义限制；词性类的概念解决了自动词性标注与查错评分标准相同的问题。⑵混淆集的概念将校对问题转化为排歧问题，使查错的目的性更强。⑶Winnow 方法非常适合校对中遇到的特征空间规模巨大、且特定的目标只和整个空间中少数特征相关的学习问题。实验表明我们的方法具有较高的查错召回率和查错准确率。

参考文献

1. 梅家驹,竺一鸣,高蕴琦,等.《同义词词林》. 上海:上海辞书出版社,1983.
2. Zhang Zhaohuang. A Pilot Study on Automatic Chinese Spelling Error Correction. Communication of COLIPS,1994,4(2):143-149
3. Sun Cai. Research on Lexical Error Detection and Correction of Chinese Text: [Master's Degree Dissertation]. Beijing: Tsinghua University Computer Science and Technology Department,1997
4. Gale W A, Church K W, Yarowsky D. A method for disambiguating word senses in a large corpus. Computers and the Humanities, 1993,26:415-439
5. Golding A R. A Bayesian hybrid method for context-sensitive spelling correction. In: Proc. 3rd Workshop on Very Large Corpora, Boston, MA:1995
6. Golding A R, Dan R. Applying Winnow to context-sensitive spelling correction. In: Proc. the 13th ICML, Bari, Italy:1996
7. Yarowsky D. Decision list for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: Proc. 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM:1994