文章编号: 1003－0077(2004)01－0026－07

中文文本分类中特征抽取方法的比较研究①

代六玲[1,2], 黄河燕[2], 陈肇雄[2]

(1. 南京理工大学 计算机科学系, 南京 210094;
2. 中国科学院 计算机语言信息工程研究中心, 北京 100083)

摘要: 本文比较研究了在中文文本分类中特征选取方法对分类效果的影响。考察了文档频率 DF、信息增益 IG、互信息 MI、χ2分布 CHI 四种不同的特征选取方法。采用支持向量机(SVM)和 KNN 两种不同的分类器以考察不同抽取方法的有效性。实验结果表明, 在英文文本分类中表现良好的特征抽取方法(IG、MI 和 CHI)在不加修正的情况下并不适合中文文本分类。文中从理论上分析了产生差异的原因, 并分析了可能的矫正方法包括采用超大规模训练语料和采用组合的特征抽取方法。最后通过实验验证组合特征抽取方法的有效性。

关键词: 计算机应用; 中文信息处理; 文本自动分类; 特征抽取; 支持向量机; KNN
中图分类号: TP18 文献标识码: A

A Comparative Study on Feature Selection in Chinese Text Categorization

DAI Liu-ling[1,2], HUANG He-yan[2], CHEN Zhao-xiong[2]
(1. Department of Computer Science, NUST, Nanjing 210094, China;
2. Language Information Engineering, CAS, Beijing 100083, China)

Abstract: This paper is a comparative study of feature selection methods in text categorization. Four methods were evaluated, including document frequency (DF), information gain (IG), mutual information (MI) and χ^2-test (CHI). A Support Vector Machine (SVM) and a k-nearest neighbor (KNN) were selected as the evaluating classifiers. We found IG, MI and CHI had poor performance in our test, though they behave well in English text categorization. We analyzed the reasons theoretically and put forwarded the possible solutions. A furthermore experiment proved that the combined feature selection method is effective.

Key words: computer application; Chinese information processing; text categorization; feature selection; SVM; KNN

1 引言

文本自动分类任务是对未知类别的文字文档进行自动处理, 判别它们所属预定义类别集中的一个或多个类别。随着各种电子形式的文本文档以指数级的速度增长, 有效的信息检索、内容管理及信息过滤等应用变得越来越重要和困难。文本自动分类是一个有效的解决办法, 已成为一项具有实用价值的关键技术。近年来, 多种统计理论和机器学习方法被用来进行文本的自动分类, 掀起了文本自动分类的研究和应用的热潮。

① 收稿日期: 2003－09－22
基金项目: 国家自然科学基金资助项目(60272088)
作者简介: 代六玲(1977－), 男, 博士研究生, 主要研究方向为中文信息处理.

文本自动分类问题的最大特点和困难之一是特征空间的高维性和文档表示向量的稀疏性。在中文文本分类中,通常采用词条作为最小的独立语义载体,原始的特征空间由可能出现在文章中的全部词条构成。而中文的词条总数有二十多万条,这样高维的特征空间对于几乎所有的分类算法来说都偏大。寻求一种有效的特征抽取方法,降低特征空间的维数,提高分类的效率和精度,成为文本自动分类中需要首先面对的重要问题。

近年来在中文文本自动分类中使用较多的特征抽取方法包括文档频率 *DF*、互信息 *MI*、信息增益 *IG* 和 χ^2 统计[1~3]等。选择特征抽取方法的一个依据是 Y. Yang 的实验[4]。由于中文与英文的文本分类问题具有相当大的差别,体现在原始特征空间的维数更大,文章表示更加稀疏,词性变化更加灵活等多个方面。在英文文本分类中表现良好的特征抽取方法未必适合中文文本分类。对中文文本分类中的特征抽取方法进行系统的比较研究十分必要。

本文的组织如下:第二部分简述本文考察的各种特征抽取方法;第三部分描述实验设置,包括文档表示方法、分类器和实验数据;第四部分给出实验结果和结果分析,并分析了可能的矫正方法。第五部分为结论。

2 特征抽取方法

2.1 文档频率

词条的文档频率(Document Frequency)是指在训练语料中出现该词条的文档数。采用 DF 作为特征抽取基于如下基本假设:*DF* 值低于某个阈值的词条是低频词,它们不含或含有较少的类别信息。将这样的词条从原始特征空间中移除,不但能够降低特征空间的维数,而且还有可能提高分类的精度。

文档频率是最简单的特征抽取技术,由于其具有相对于训练语料规模的线性计算复杂度,它能够容易地被用于大规模语料统计。但是在信息抽取(Information Retrieval)研究中却通常认为 *DF* 值低的词条相对于 *DF* 值高的词条具有较多的信息量,不应该将它们完全移除。Y. Yang 的实验证明:在英文环境中,当 *IG* 和 *CHI* 等统计方法的计算"费用"太高而变得不可用时,DF 可以安全的代替它们被使用。我们将在中文环境中重新检验 DF 的有效性。

2.2 信息增益

信息增益(Information Gain)在机器学习领域被广泛使用[5]。对于词条 t 和文档类别 c,*IG* 考察 c 中出现和不出现 t 的文档频数来衡量 t 对于 c 的信息增益。我们采用如下的定义式:

$$IG(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \quad (1)$$

其中 $P(c_i)$ 表示 c_i 类文档在语料中出现的概率,$P(t)$ 表示语料中包含词条 t 的文档的概率,$P(c_i \mid t)$ 表示文档包含词条 t 时属于 c_i 类的条件概率,$P(\bar{t})$ 表示语料中不包含词条 t 的文档的概率,$P(c_i \mid \bar{t})$ 表示文档不包含词条 t 时属于 c_i 的条件概率,m 表示类别数。

实验中我们对在语料中出现的每个词条计算其信息增益值,从原始特征空间中移除低于特定阈值的词条,保留高于阈值的词条作为表示文档的特征。

2.3 CHI 统计

CHI 统计方法度量词条 t 和文档类别 c 之间的相关程度,并假设 t 和 c 之间符合具有一阶自由度的 χ^2 分布[6]。词条对于某类的 χ^2 统计值越高,它与该类之间的相关性越大,携带的类别信息也较多。令 N 表示训练语料中的文档总数,c 为某一特定类别,t 表示特定的词条,

A 表示属于 c 类且包含 t 的文档频数，B 表示不属于 c 类但是包含 t 的文档频数，C 表示属于 c 类但是不包含 t 的文档频数，D 是既不属于 c 也不包含 t 的文档频数。则 t 对于 c 的 *CHI* 值由下式计算：

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B) + (C + D)} \tag{2}$$

对于多类问题，分别计算 t 对于每个类别的 *CHI* 值，再用下式计算词条 t 对于整个语料的 *CHI* 值，分别进行检验：

$$\chi^2_{\max}(t) = \max_{i=1}^{m} \chi^2(t, c_i) \tag{3}$$

其中 m 为类别数。从原始特征空间中移除低于特定阈值的词条，保留高于该阈值的词条作为文档表示的特征。另一种方法是将词条对于各个类别的平均 *CHI* 值作为它对所有类别的 *CHI* 值，但是它的表现不如(3)式。

2.4 互信息

互信息(Mutual Information)在统计语言模型中被广泛采用[7]。如果用 A 表示包含词条 t 且属于类别 c 的文档频数，B 为包含 t 但是不属于 c 的文档频数，C 表示属于 c 但是不包含 t 的文档频数，N 表示语料中文档总数，t 和 c 的互信息可由下式计算：

$$MI(t, c) \approx \log \frac{A \times N}{(A + C) \times (A + B)} \tag{4}$$

如果 t 和 c 无关(即 $P(tc) = P(t) \times P(c)$)，$I(t, c)$ 值自然为零。为了将互信息应用于多个类别，与 *CHI* 统计的处理类似，由下式计算 t 对于 c 的互信息：

$$MI_{\max}(t) = \max_{i=1}^{m} I(t, c_i) \tag{5}$$

其中 m 为类别数。将低于特定阈值的词条从原始特征空间中移除，降低特征空间的维数，保留高于阈值的词条。

3 实验设置

实验分为建立特征库、分类模型训练和分类测试三个步骤。在对文档进行进一步处理之前需要对它进行预处理，包括分词和停用词的移除。分词词典的规模为 182966 条。停用词是指不包含类别信息的词汇，如：的、是、总之等。在实验中使用的停用词总数为 1094 条。

分别统计每个词条的 *DF*、*MI*、*IG* 和 *CHI* 值，设定合适的阈值，将特征值低于该阈值的词条移除，构成不同大小的特征库。在训练和分类模块中，依据特征库对文本进行特征提取，进而将文档表示为特征向量(见 3.1 节)。训练模块生成分类模型，分类模块根据分类模型对测试文本的类别做出预测。

3.1 文档表示

对文档进行分类之前需要将文档表示为计算机能够处理的形式。向量空间模型(VSM)是使用较多且效果较好的表示方法之一[8]，在该模型中，文档空间被看作是由一组正交向量张成的向量空间。若该空间的维数为 n，则每个文档 d 可被表示为一个实例特征向量 $V(d) = (\omega_1, \omega_2, \cdots, \omega_n)$，$V$ 的每一个分量表示对应特征在该篇文档中的权值。

计算特征权值 ω 的一种方法是 *TFIDF*[9]。词条 t_i 在文档 d 中的 *TFIDF* 值由下式定义：

$$TFIDF_i = TF_i \times \log(N / DF_i) \tag{6}$$

其中 TF_i 是词条 t_i 在文档 d 中出现的频数，N 表示全部训练文档的总数，DF_i 表示包含词条 t_i 的文档频数。为降低高频特征对低频特征的过分抑制，在实验中计算权值时还对

TFIDF 值进行 L^2 规范化处理：

$$\omega_i = \frac{TFIDF_i}{\sqrt{\sum_{j=1}^{n}(TFIDF_j)^2}} \tag{7}$$

3.2 分类器

为了消除特征抽取方法对分类器可能的倚重，在实验中我们分别使用 *KNN*（*K* Nearest Neighbor）和支持向量机（Support Vector Machines）作为文本分类器。

KNN 是一种传统的模式识别方法[10]，被广泛的应用于文本自动分类研究[11]，在准确率和召回率上表现出众。*KNN* 在首先在已知类别样本中寻找与待分类样本 X 最相似的 K 个样本，文本样本之间的相似性可以通过文本向量之间的余弦来度量：

$$Sim(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\| \times \|\vec{Y}\|} \tag{8}$$

KNN 基于这 K 个已知类别样本的类别属性对未知样本的类别做出预测。一种简单的预测规则就是将未知样本的类别预测为在这 K 个最近邻样本中包含最多实例的类别。即：

$$C(X) = \arg\max_i \{ n(d_j, c_i) \mid j \in [1, K] \} \tag{9}$$

式中 K 表示 X 的最近邻的个数，$n(d_j, c_i)$ 表示最近邻中属于 c_i 类别的样本个数。我们分别令 K 取 3 到 29 范围内的不同奇数值进行测试，并将最优的结果用于比较。

SVM 由 VapniK 发明[12]，是一种相对较新的机器学习技术，近年已被广泛地用于模式识别的多个领域[13]，取得了非常好的效果。通过学习算法，*SVM* 在训练样本中寻找具有最好区分能力的样本点集，称为支持向量（Support Vectors）。在分类阶段，*SVM* 利用这些支持向量对未知类别样本的类别属性做出预测。我们基于 C/C++ 实现了改进的 *SVM* 训练算法 *SMO*[14, 15] 用于实验。实验中核函数采用高斯函数，乘子上界 C 取 1.5，σ^2 取 1。

3.3 数据集

实验中使用 TREC-5 人民日报新闻语料库。TREC-5 新闻语料于 1995 年由 LDC（Linguistic Data Consortium）发布。其中包含国内、国际、体育、文化等新闻文章共 77733 篇，每一篇文章都包含 SGML 格式的标题，指示该篇文章的类别。根据标题内容，我们从 TREC-5 中抽取出政治社会、经济、文化教育、文艺娱乐、体育、学术理论六类，每类 1000 篇共 6000 篇文章用于实验。经过分词并移除停用词之后，共有不同词条 49305 条。

4 实验结果及分析

4.1 性能评价

特征抽取作为分类的前处理过程，其有效性可以通过分类的效果来测试。为评价分类效果，我们采用最通用的性能评价方法：召回率 R（*Recall*）、准确率 P（*Precision*）和 F_1 评价。对于某一特定的类别，召回率定义为被正确分类的文档数和被测试文档总数的比率，即该类样本被分类器正确识别的概率。准确率定义为正确分类的文档数与被分类器识别为该类的文档数的比率，即分类器做出的决策是正确的概率。通常还将召回率和准确率用某种方式组合成单一的度量，以便于进行比较。我们使用 F_1 度量这种较通用的组合方式：

$$F_1 = \frac{2RP}{R+P} \tag{10}$$

在实验中将分类器设计为两类的，分别对实验语料中的每一类进行实验。选定某一类样本为正样本，则其余各类都是负样本。在分类模型训练阶段，从正样本中随机抽取 800 篇作为

正训练样本，从其余各类中平均随机抽取 1600 篇作为负训练样本。分类阶段中使用剩余的 200 篇正样本作为正测试样本，从其余各类中平均随机抽取 400 篇作为负测试样本，并且保证测试集和训练集没有交集。记录每个类别的 F_1 值之后，再计算它们的广义均值(Micro Average)，最后使用广义均值对特征抽取方法的有效性给出评价。

4.2 实验结果

图 1 和图 2 分别所示在采用不同的特征抽取方法之后，*SVM* 和 *KNN* 在 TREC-5 语料上的 *F*1 均值曲线。在我们的实验中特征空间的维数上限取 30000(在实际设计分类器时，特征空间的维数一般不会超过一万)。

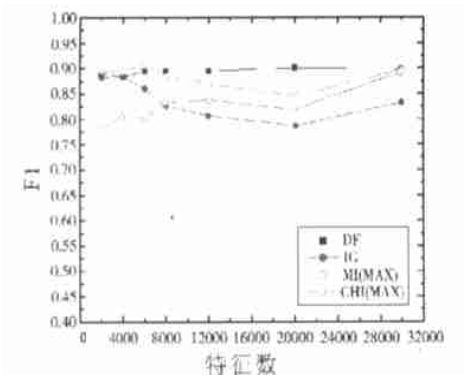

图 1 特征抽取方法在 SVM 上的表现

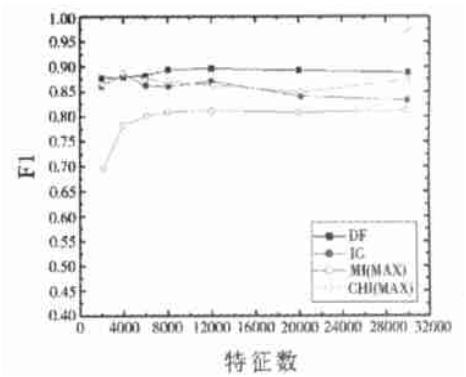

图 2 特征抽取方法在 KNN 上的表现

从 *SVM* 和 *KNN* 的 F_1 曲线中可以看出：在 *SVM* 和 *KNN* 上，*DF* 的表现都优于其它方法。*MI* 在则表现得不稳定，在 *SVM* 上，当特征空间的维数大于 8000 时，其效果超过 *IG*，而在 *KNN* 中却表现得最差。另外，当特征空间取值在 5000 到 10000 范围上时，分类器的性能都达到一个极大值。这意味着可以将特征空间的维数压缩到原始特征空间维数的 10%～20%，移除其余 80%～90%的特征。

4.3 结果分析

虽然 *IG*、*CHI* 和 *MI* 在英文文本分类问题中表现良好，但是在我们的实验中它们的表现远远不及 DF。经过仔细分析发现，造成这种差别的原因来自于两方面：使用类别信息的特征抽取方法对低频词的倚重和中文相对于英文具有更高的特征空间维数。

IG、*CHI* 和 *MI* 都通过不同的方式使用词条的类别信息。具体来说，*IG* 计算 $P(c_i|t)$ 和 $P(c_i|\bar{t})$ 的值，并同类别的概率一起度量词条携带的类别信息；*CHI* 的定义式中的 *A*、*B*、*C* 和 *D* 都表示在训练文档中词条的类别信息；对于 *MI*，将它的定义式(4)稍加变形可得等价式：$MI(t,c)=\log P(t|c)-\log P(t)$，其中 $\log P(t|c)$ 即为类条件概率。

当训练语料库的规模没有达到一定规模的时候，特征空间中必然存在相当数量的出现频数很低(比如低于三次)甚至不出现的特征。而因为它们较低的出现频数，必然只属于少数的类别。而使用类别信息的统计方法必定认为这些低频词携带较为强烈类别信息，从而对它们有不同程度的倚重。但是经过仔细观察发现，这些低频词中只有不到 20%的词确实带有较强的类别信息，大多数的词都是噪音词，不应该成为特征。

在英文文本的分类问题中，通常取单词和短语作为特征，特征空间的维数相对较少。例如 Y. Yang 的实验中特征空间的维数为 16000。在训练语料的规模适度大(这样的规模教容易达到)的情况下，大多数词条都可以获得较高的 *DF* 值，使得 *IG*、*CHI* 和 *MI* 对低频词的倚重

得到减弱。但是，正是由于这种倚重的存在，使得在 *Y . Yang* 的实验中它们的表现没能明显优于 *DF*，甚至不如 *DF*。

在中文文本中分类应用中，通常将单个的词条作为特征。如果不考虑人名和地名等未登陆词，可以近似认为特征空间维数等于分词词典中的词条数目。在中文处理中，通常采用的分词词典的规模一般在 5 万到 25 万词条之间[16]。也就是说中文的特征空间维数比英文更高，而且可能高很多。在相同规模训练语料条件下，更高的维数必然导致更多的低频词出现。在这样的情况下使用 *IG*、*CHI* 和 *MI* 进行特征抽取，由于它们对低频词的倚重，必定会将更多的低频词作为特征使用。从而导致了分类效果的低下。

4.4 矫正方法分析

基于以上的分析可知，要使用类别信息提高分类效果，必须消除相应的特征抽取方法对低频词的倚重。一个方法是增加训练语料的规模，使得所有或至少绝大多数词条在语料中的出现频数都超过一定阈值。但是由于中文特征空间的维数如此之高，要达到该目的训练语料至少要达到 GB 级。在实际设计文本分类器的时候，如此大规模的训练语料通常难于获取。再者 *IG*、*CHI* 和 *MI* 的计算复杂程度都为 $O(N^2)$，统计如此大规模的语料将会花费很高的计算成本。

在只有通常规模训练语料的条件下，充分利用类别信息的一个可行方法是使用组合的特征抽取方法，即使用 *DF* 与 *IG*、*CHI* 或 *MI* 结合，用于特征抽取。先使用 *DF* 移除低于一定阈值的低频词，消除 *IG*、*CHI* 或 *MI* 对低频词的倚重。再使用 *IG*、*CHI* 或 *MI* 从剩余词条中移除类别信息较低的噪音词，这样的词条通常平均分布于各个类别。由于 *MI*、*IG* 和 *CHI* 本身的计算中都利用到词条的 *DF* 值，组合特征抽取方法并不会增加计算量。

作为验证，我们分别将 *DF*＋*IG*、*DF*＋*CHI* 和 *DF*＋*MI* 用于 *SVM*，*DF* 的阈值设定为 3，图 3 给出了实验结果。从图 3 可以看出，在我取方法的对分类效果的提高是显著的。另外，实验发现，采用组合的特征们的语料上，组合的特征抽抽取方法之后，*SVM* 分类器的训练时间被大大缩短。图 4 分别示出采用 *DF*＋*IG*、*DF*＋*MI*、*DF*＋*CHI* 后 *SVM* 的训练时间与采用 *DF* 时的比较，组合特征抽取方法使得 *SVM* 的训练时间缩短了 5—13 倍。训练时间的缩短从另一侧面说明了组合方法是有效的。

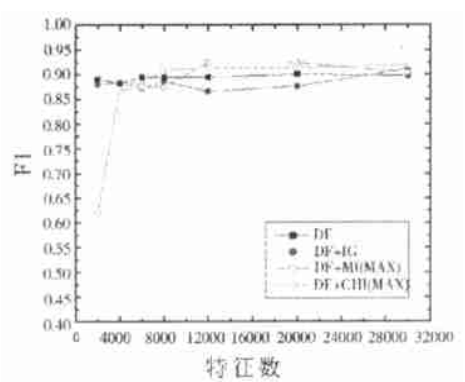

图 3　组合特征抽取方法对文本分类效果的影响

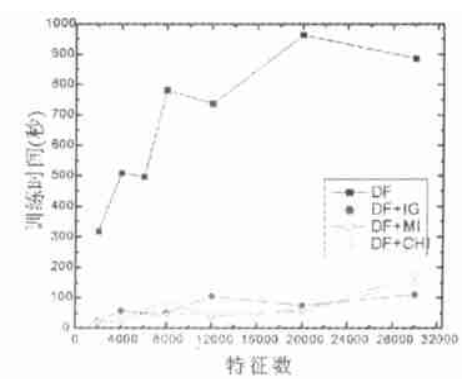

图 4　组合特征抽取方法对 SVM 训练时间的影响

5　结论

本文考察了在中文文本分类中常用的四种特征抽取方法：文档频率 *DF*、信息增益 *IG*、

CHI 统计和互信息 *MI*。文中发现利用类别信息的特征抽取方法在不加以修正的情况下并不适合中文文本分类这一事实。并分析了产生这种现象的原因在于利用类别信息的特征抽取方法对低频词的倚重和中文的特征空间维数远远高于英文特征空间维数两个方面。文中分析了可能的矫正措施包括增大训练语料的规模和采用组合的特征抽取方法。在大多数情况下,后者更为实用。进而的实验结果表明组合的特征提取方法不但提高了分类的精度,同时还显著的缩短了分类器的训练时间。本文的贡献在于为设计中文文本分类器时特征抽取方法的选择提供了指导,相关结论同样适合其它语种。

参 考 文 献:

[1] 孙丽华, 等. 一种改进的 KNN 方法及其在文本分类中的应用[J]. 应用科技第 29 卷第 2 期2002 年2月

[2] 朱寰, 等. 文本分割算法对中文信息过滤影响研究[J]. 计算机工程与应用, 第 13 期, 2002

[3] 何新贵, 等. 中文文本的关键词自动抽取和模糊分类[J]. 中文信息学报, 1999: 13(1)

[4] Y. Yang. A Comparative Study on Feature Selection in Text Categorization[C]. In: Proceeding of the Fourteenth International Conference on Machine Learning (ICML'97), 412—420, 1997.

[5] Tom Mitchell. Machine Learning[M]. McCraw Hill, 1996.

[6] T. E. Dunning. Accurate methods or the statistics of surprise and coincidence[C]. In: Computational Linguistics, Volume 19: 1, pages 61—74, 1993.

[7] Kenneth Ward Church and PatricK Hanks. Word association norms, mutual information and lexicography[C]. In: Proceedings of ACL27, pages 76—83, Vancouver, Canada, 1989.

[8] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[C]. Communications of th ACM, 1975, 18(5): 613—620.

[9] Salton G. Introduction to Modern Information Retrieval[M]. New York: McGraw-Hill Book Company, 1983.

[10] Belur V. Dasarathy. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques[M]. IEEE Computer Society Press, Las Alamitos, California, 1991.

[11] Y. Yang and X. Liu. A re-examination of text categorization methods[C]. In: Proceedings, 22nd Annual International ACM SIGIR Conference on Research and Develop-ment in Information Retrieval (SIGIR'99), 42—49, 1999.

[12] Vladmimir N. VapniK. The Nature of Statistical Learning Theory[M]. Springer, New York, 1998.

[13] Burges, C. J. C. A tutorial on support vector machines for pattern recognition[C]. Data Mining and Knowledge Discovery, 1998, 2(2): 955—974.

[14] J. Platt. Fast training of support vector machines using sequential minimal optimization[C]. In: B. ScholKopf, C. Burges, and A. Smola, editors, Advances in Kernel methods: support vector learning. MIT Press, 1998.

[15] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvementsto platt's SMO algorithm for SVM classifier design[R]. Neural Computation, 13(3): 637—649, March 2001.

[16] 黄昌宁, 等. 对自动分词的反思[C], 语言计算与基于内容的文本处理, 北京: 清华大学出版社. 26—38, 2003, 7.