文章编号：

基于转换表及上下文环境的汉语简繁文本双向翻译 *

庞祯军 姚天昉

（上海交通大学计算机科学与工程系，上海 200240）

摘要：现有的简繁转换技术在处理简繁一对多时效果不是很理想。为了解决这一问题，作者提出了基于转换表和上下文的汉语简繁文本双向翻译方法。作者之前的研究工作成果在教育部语信司所举行的简繁一对多转换评测中取得了 95.6%的转换准确率。在此研究基础上，本文提出了使用规则加组合统计模型来解决这一问题，所组合的统计模型为 SVM、最大熵和 Bayes 模型。同时作者还提出了一种提高文本分类准确度的新的特征选择方法 ADMMR，该方法和 ECE，卡方检验这两种特征选择方法具有相当的性能；同时还提出了最大熵模型的特征值使用 tf-idf，而不使用 0-1 值。实验表明这一调整使准确度提高了约 2%。此外，作者使用 ADMMR、ECE 和卡方检验作为文本的特征选择方法，使用 tf-idf 来量化每一个特征，经过实验表明组合模型在处理一简对多繁问题时具有更高的转换准确率和更稳定的性能。实验表明规则加组合模型的方法能够达到 98.5%的准确率，较好地解决了简繁转换中的一对多转换的问题。

关键词：简繁转换，简繁一对多转换，组合模型，SVM，最大熵，GIS，ADMMR，特征选择

中图分类号：TP391 **文献标识码**：A

Chinese Bilateral Translation between Simplified and Complex-Character Texts based on Conversion Table and Context

PANG Zhen-jun, YAO Tian-fang

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240)

Abstract: The existing Simplified and Traditional Chinese text translation technologies are not very ideal.In order to solve this problem, this paper proposes a method thatcombinesrules and models.The authors' previousworkhada conversion accuracy rate at 95.6% in a test which was done by the Ministry of Education.On the basis of this study, the paper presents a combination rules and models which includeSVM, maximum entropy model and Bayes model. At the same time, the authors propose a method named ADMMR to improve the accuracy of feature selection,and the methodhas a great performance. Moreover, the authors use ADMMR, ECE and the Chi-Square test of text as feature selection methods, and then useTF-IDF to quantify each feature. The experiments showed that the combined modelsin dealing with more complicated issue have higher and morestable performance, that is to say, the combination of rules and modelsachieves 98.5% accuracy rate. Therefore,it is a much better solution to the conversion problem.

Keywords: Simplified and Traditionaltext conversion,Simplified and Traditional one-to-many conversion, combination model, maximum entropy, SVM, GIS, ADMMR, feature selection

1 引言

海峡两岸都是使用汉字，但是因为政治和历史的原因，两岸双方现在使用的汉字在字形和用词等方面都出现了很大差异，现行的汉字分为简体字和繁体字，其中中国大陆和新加坡等地区使用简体字，港澳台地区和部分海外华人使用繁体字[1]。两岸四地同源同种交流密切，特别是在 2008 年台湾政党轮替之后大陆地区和台湾地区的交流更加密切，简体字和繁体字

*收稿日期：2015-08-04 定稿日期：2015-08-08

作者简介：庞祯军（1987—），男，硕士，主要研究方向为意见抽取，信息抽取，自然语言处理，pzj_636484@163.com；姚天昉（1957—），男，博士，副教授，硕导，主要研究方向为意见挖掘、信息抽取、机器学习、自然语言处理等，yao-tf@cs.sjtu.edu.cn。

的差异给交流带来了文字障碍，在这种背景下，两岸之间的不少有志之士呼吁应互相协商，统一两岸文字，消除交流障碍达到“书同文”的状态。海峡两岸之间文字的最大差异是大陆地区使用简化字，台湾地区使用繁体字，为了消除交流障碍，有不少专家学者建议先在学术层面上讨论汉字统一的标准，在此之前我们需要先弄清楚简化字和繁体字之间的各种差异，找出简化字和繁体字之间的对应关系 [2-6]，以便于中文信息处理和交流，为今后“书同文”打下基础。因此消除这些文字障碍对交流有着极大的促进作用，对两岸经济和文化交流、合作和发展有着深刻的意义，这也使得研究简繁文本的翻译有着重大的现实意义 [7]。

现行简化字以 1964 年公告，1986 年修订的《简化字总表》为国家标准，《简化字总表》共收 2274 个简体字。一般情况下，一个简体字是和一个繁体字一一对应的，但有部分简体字是对应多个繁体字的，同时一个繁体字也可能对应多个简体字，这种一对多的关系是简繁汉字翻译的难点；同时随着语言的自然发展而产生了一些两岸人民各自的惯用语及专业术语，这也是简繁文本翻译的一个难点。

简繁双向翻译涉及到的一对多问题和术语对应为题可以在这个例子中很明显看出，对于简体文本“NBA 篮网队与开拓者队的比赛中，篮网队受困于无暂停可用，总教练基德授意球员撞掉其手中的咖啡弄湿地板，裁判只能暂停比赛叫工作人员擦干地板，篮网队趁机布置了战术并将比赛拖入加时赛。基德干的漂亮！”其所对应的繁体文本为“NBA 籃網隊與拓荒者隊的比賽中，籃網隊受困於無暫停可用，總教練基德授意球員撞掉其手中的咖啡弄濕地板，裁判只能暫停比賽叫工作人員擦乾地板，籃網隊趁機佈置了戰術並將比賽拖入延長賽。基德幹的漂亮！”存在一对多关系的简体字有“只”，“干”，“了”等，两岸不同的术语则有：“开拓者队”对应“拓荒者隊”；“加时赛”对应“延長賽”。

简化字和繁体字之间的对应关系从语义层次上来看基本是明确的，只有部分词是例外，因此在转换时只需要先考虑这些例外，再根据固定用法和语义上下文来判断简化字是对应到哪个繁体字，这就是本文处理简化字和繁体字转换的基本思路。

2 相关研究

本文所研究的课题有不少计算机方面的学者专家进行过研究，但因年代限制其所处时代机器学习方法还未普遍流行，因此只能采用规则处理，但是规则处理无法较好地处理单字不成词时的简繁转换[8-12]，因此在简繁转换这一功能上未能有较大的突破。刘汇丹[13]等分析了简繁汉字转换的复杂性，并对字符编码集进行了分析。GB2312-80只收录了6763个简体中文常用字和次常用字，TCA-CNS11643-1992收录汉字13053个，但是这两个字符集都没有包含所有的简体字和繁体字，因此简繁转换的过程中需要对编码集进行处理，相对来说就增加了简繁汉字转换的难度。而国际标准编码字符集Unicode/ISO-IEC10646（以下简称Unicode）为世界上所有的文字进行统一的编码，在此编码集下简繁转换就无需考虑编码集的问题。作者们在分析了简繁转换的相关问题后，对系统进行了实现，整个系统所使用到的技术分为：分层次转换、词语消歧、命名实体识别、转换正确性评估、分词和词典查找等；并根据转换思想例举了一个转换的简单示例，不过因为作者所收集的资料不是很充分，所以未能给出系统的转换准确率数据。从整体上来说，该文所述系统基本涵盖了简繁转换所需处理的问题，但是其未能解决简繁转换中的最难点即单字无法成词时消歧的处理。辛春生[14]等构建了一个使用规则方法来处理简繁汉字转换的系统，其主要思路是通过构建一个很大的词语库来对完成简繁转换过程，对于一对多的简体字的处理也是按照这个思路。在词语切分过程中可能产生的切分歧义则分为三个步骤来处理：根据系统定义的规则和用户定义的规则来消歧；根据语法和语义分析来消歧；根据词频来消歧。在经过消歧之后就能够得到相应的词语链，再将词语链中的词语转换到目标词。但作者的这一思路无法攻克简繁转换中的难点问题，即一对多简化字的简繁转换特别是当该简化字无法成词时的简繁转换。

统计模型在简繁转换中有着很重要的作用，不仅仅在切分词语时需要使用统计模型，而且在判别简化字的转换目标时也需要统计模型。辛春生[15]等研发了一套简繁转换系统，并且深入研究了词语切分算法对于简繁转换准确率的影响。该文作者从《人民日报》随机抽取了19540的语料测试该简繁转换系统的准确度可以达到99.93%，又从《中国计算机报》随机抽取了16051字的语料进行简繁转换发现可以达到的转换准确率为99.85%。该文所存在的主要问题是认为一对多简繁转换的解决方法为增加简繁转换的词库，但事实上词库的方法是不可能很好地解决这一问题的。同时作者的相关测试语料太少，不能准确反映系统的实际性能。该文中所介绍的汉字编码之间的转换和计算机内码之间的转换是一个比较有特点的，简繁转换中的其他文章基本只是介绍相关知识而没有做到实际的转换，不过现有的Unicode编码也是可以完美地解决简繁转换中的汉字编码问题。

对于简繁转换的难点一对多简体字转繁体字的问题，也有学者尝试采用统计模型[16]的方法去解决。李民祥[17]等设计了一个具有扩展性的简繁转换系统，并且采集了维基百科上的术语对照表词条，而对于一对多的简繁转换问题则采用语言模型来处理。例如对于简体句子“剪发之后，感觉整个人都清爽多了！”，对应的繁体文本可为“剪（髮發）之後，感覺整個人都清爽多了！”，对于句子中的简体字“发”字计算“剪髮”、“剪發”、“發之”和“髮之”的bigram值，选择bigram值较高的情况作为转换的结果。这里的bigram也可以换为trigram或者其他的N-gram，经过作者分析，当N越大时转换的准确率越高。作者根据N-gram来判断需转换到的目标词有一定的合理性，但是N-gram越大不一定是正确的选择，因为这一原因使得该系统会产生断词错误从而导致转换错误，在这分析了这些转换错误之后，作者引入了中科院的分词系统并使得系统的转换准确率提高了1%。因此，该文使用统计模型来判别转换结果的方法有一定的参考意义。

统计模型在解决分类问题上有其独特的优势，特别是对于涉及上下文信息的分类问题。最大熵模型也常用于解决文本分类问题，因此也有学者使用了最大熵模型来处理简繁转换。Fai Wong[18]等为了避免构造词库所耗费的人力物力而采用了最大熵模型，对于特征选择则采用N-gram。经过他们的测试，该简繁转换系统在处理一对多简繁汉字转换时的准确度能够达到89.94%，比微软的office所具有的简繁转换准确度87.86%略高，具有一定的积极意义。但是对于一对多简繁转换89.94%还远远无法达到实用的要求，经过分析最大熵模型用于文本分类时的效果，可以发现特征选择在其中起到了一个非常重要的作用，而Fai Wong等采用的N-gram特征选择方法相比较其他特征选择方法不具有比较优势，因此使得最终的转换准确率不是相当理想。

厦门大学信息技术系和认知科学系的相关研究人员在简繁转换这一课题上有比较深入的研究[19-20]。Yidong Chen[21]等对于简繁转换的一对多问题，提出使用log-liner 模型，在这个模型中使用的特征函数有：词汇语义一致性权重、语言模型特征、句子长度惩罚和词组数量惩罚。在考虑词汇语义一致性权重的同时也考虑了交叉语言信息对系统的影响，实验表明，Yidong Chen等所提出的方法简繁转换效果显著。Xiaodong Shi[22]等则在Yidong Chen所提出方法的基础上做了优化，第一，对于所使用的相关数据进行了去噪声处理；第二，增加了贝叶斯模型作为log-liner框架模型中的一个翻译模型，同时在特征选择时采用了互信息作为特征选择的方法。通过作者的工作显示出简繁转换的相关语料的准确度的确需要加以过滤，特别是如果采用新闻语料则要更加谨慎地处理。

3 简繁转换流程及模型

对于其中的一对多简体字转繁体字问题，本文提出使用规则加组合模型的方法来解决：规则方法即是首先将待转换的文本进行分词，根据词库将简体词组转换到对应的繁体词组；若规则无法处理则使用分类模型把简体字转换到对应的繁体字，即此时将该问题看成一个文本分类问题。通过使用统计方法来学习该简体字所在的上下文环境从而得到分类模型，再应

用分类模型对待转换的简体字进行处理。本文所使用的词库数据是从《臺灣國語辭典》和维基百科[23]中收集到的；模型的训练数据则部分为维基百科简繁对应文本和简繁体小说。

3.1 简繁汉字转换流程

简繁汉字转换分为简体转繁体，繁体转简体以及包含其中的术语转换。在此分为简体转繁体以及繁体转简体进行转换。

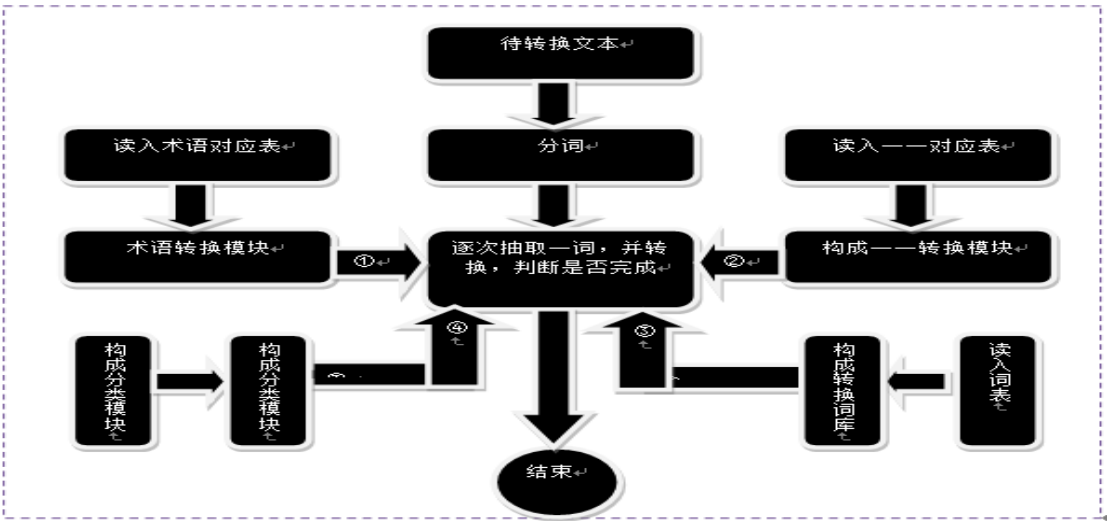


图 1 简体字转繁体字流程

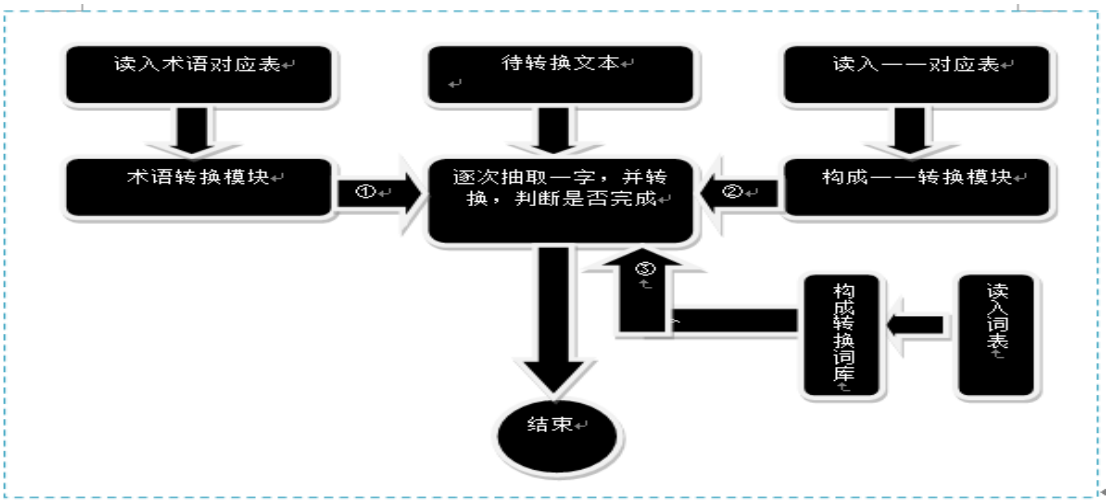


图 2 繁体字转简体字流程

3.2 简繁汉字转换模型

简繁汉字转换的重点和难点在于一个简体字对应多个繁体字的情况，一般情况下这些简体字可以和其他字成词，因此我们可以通过词语的固定用法将简体字转换到对应的繁体字，此种情况只需要收集简体词语和繁体词语的对应关系即可得到转换结果；而对于不能成词的情况，本文提出使用机器学习的方法来学习简体字所对应上下文的环境，在有足够丰富的学习语料前提下将能够很好地解决这一问题。

使用机器学习方法对文本分类涉及到如下几个步骤：

1) 文本表示方法，即如何将文本进行表示以用于模型学习。现在比较流行的是向量表示法，也有概率模型表示法。在使用这些表示法之前一般都需要对文本进行切分处理，如分词算法是将文本分为词的序列，而 N-gram 则是将文本进行特定的 N-gram 处理。如果直接

将文本进行切分之后所产生的全部词组作为特征，则会加大训练模型的空间复杂度和时间复杂度，因此在进行切分之后还需要进行特征选择以加快训练速度。在选择好特征之后，还需要对每个训练实例进行特征赋值，在不同统计模型中对特征的赋值方法也不一样，如有 tf 赋值法，tf-idf 赋值法以及 0-1 赋值法等等。最后就将每一个文本都表示成了特定的格式。

2) 将文本进行表示后，就可以使用机器学习方法对数据进行学习并在学习完成时将分类模型保存好以便分类时使用。现在比较流行的文本分类方法有 SVM（支持向量机）、最大熵分类模型以及贝叶斯分类模型等。

3) 最后就是测试步骤，通过调用第二步所生成的分类模型对测试数据进行测试，通过分类结果则可以得到不同分类模型的分类准确率和召回率。

3.2.1一对多简繁分类组合模型

对于不能使用规则进行处理也就是不能成词的或者成词后依然无法转换的简繁一对多情况，本文提出将 SVM 模型、最大熵模型和贝叶斯模型应用于简繁分类中的一对多问题，以达到可以解决一对多简体字无法成词时根据上下文来确定目标这一目的。本文作者在解决这一问题的过程中，提出了一种新的特征选择方法来提高文本表示的准确度；同时提出最大熵分类模型的每个特征值使用 tf-idf，而不使用 0-1 值；作者选择了 ADMMR，期望交叉熵以及卡方检验这 3 种特征选择的方法来表示文本，再使用 SVM 和最大熵模型来学习这三种文本数据从而产生 3×2=6 个分类模型，再根据贝叶斯模型学习训练数据产生第 7 个分类模型；前 6 个模型采用投票的方式，哪个类别所获得的票数最多即为最后的结果。对于平局的情况如果采用随机方法去选择则只有最高不超过 50%的正确率，作者因此引进了训练和分类都比较简单并且分类效果的准确率远大于 50%的 Bayes 模型；当平局时贝叶斯模型的结果如果也为最高票则以贝叶斯模型的结果为准，否则在最高票中随机选择。经过实验表明这 7 个模型在分类时可以弥补各分类模型的不足，产生更好更稳定的分类效果，有效地解决了简繁转换的一对多问题。

组合模型的生成框图如下：

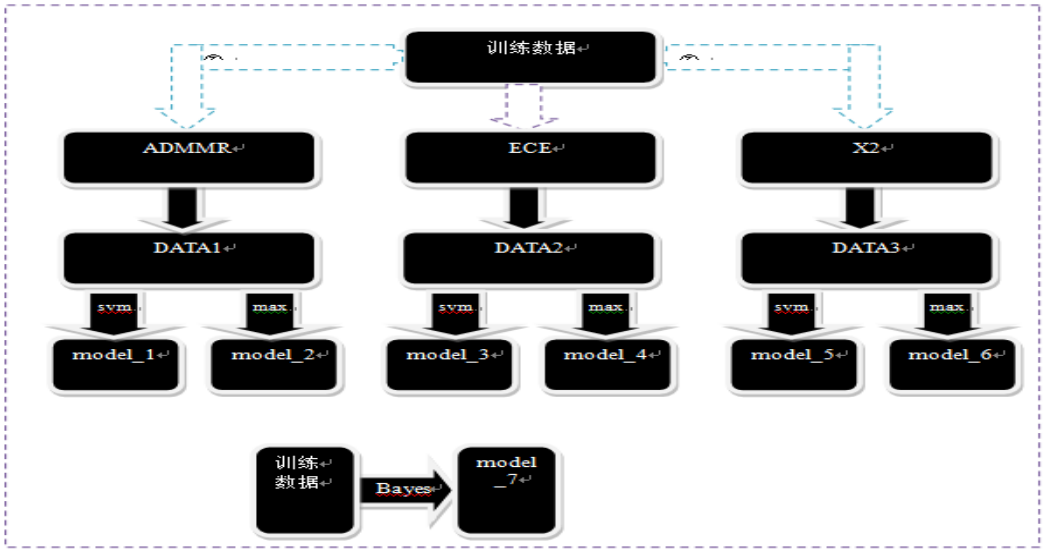


图 3 组合模型生成框图

3.3 分类模型

3. 3. 1 SVM 分类模型

支持向量机（SVM）是由超平面定义的判别分类器，也就是通过对给定训练数据的学习，SVM 可以学习到可用于分类新数据的最优超平面。

对于一个二维平面的二分类问题，找到最优的分割直线。在线性不可分的情况下，支持向量机先在低维空间下运算，将低维空间映射到高维空间，最终在高维特征空间构造出最优分离超平面，从而解决低维空间不可分的问题。如下图所示示例，这些数据在二维空间线性不可分，在映射到三维空间后，从而线性可分。

3.3.2 最大熵分类模型

熵最初是热力学领域的概念，香农在 19 世纪 40 年代在信息论中引入了信息熵的概念。熵是不确定性的度量，不确定性越大熵值越大。当一个随机变量服从均匀分布时，熵最大；当随机变量为常数时，熵为 0. 在信息论中，使用熵这一概念来定义信息量。对于一个变量 X，它的可能取值为$\mathrm{x}_1, x_2, \ldots, x_n$，且取这些值的概率分别为：$\mathrm{p}_1, p_2, \ldots, p_n$，那么变量 X 的熵为：

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

从熵的定义可以看出，一个变量其可能的取值越多且取每个值的概率变化越多，那么这个变量的熵就越大。

3.3.3 Bayes 分类模型

对于 n 个事件$\mathrm{A}_1, A_2, \ldots, A_n$，事件$\mathrm{A}_i$发生导致事件 B 发生的概率称为条件概率，记为$P(B|A_i)$。现在若事件 B 发生了，则是由事件$\mathrm{A}_i$发生所引起 B 发生的概率有多大？此概率可表示为：$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{k=1}^{n} P(A_k)P(B|A_k)}$，此公式即为 Bayes 定理。

3.4 特征选择算法

文本分类问题一般分为模型训练、应用模型进行分类两个阶段；在模型训练阶段则涉及到文本分词、特征选择及特征降维、文本向量表示、模型训练等步骤。特征选择和特征降维步骤可以将每个文本的向量维数降低，从而降低所有训练数据的复杂度，对训练模型所需要的时间有一个很大的改进。并且一个好的特征选择算法所选择出来的特征组合在提高分类准确度上都有很大的帮助。比较常用的特征选择算法有信息增益法(Information Gain) 、互信息(Mutual Information)、文档频率(DF)、卡方检验(CHI)、文本证据权(WET)、期望交叉熵(CE)以及几率比(OR)等等。

信息增益法(Information Gain)：在信息增益特征选择算法中，衡量特征的重要性是看这个特征可以给分类系统带来多少信息，带来信息越多的特征越重要。在文本分类问题 C 中，文本可以分成$\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{n}$共 n 个类别，且每个类别对应的概率为：$\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_\mathrm{n}$，则熵$\mathrm{H(C)} = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P_i} \log \mathrm{P_i}$。而衡量一个特征对分类系统的重要性则是看这个特征对分类系统的条件熵 H(C|X)，即对于特征 X，其可能取值为：$(\mathrm{x}_1, x_2, \ldots, x_n)$，那么$P(C|X) = \sum_{i=1}^{n} P(X = x_i)H(C|X = x_i)$。在文本分类中特征一般为词 T，而 T 可能的取值只有两种：出现和不出现。那么条件熵$H(C|T) = P(t)H(C|T = t) + P(\bar{t})H(C|T = \bar{t})$。

期望交叉熵又称为 KL 距离，反映的是文本类别的分布和以某特征出现前提下文本类别分布之间的距离，特征 T 的期望交叉熵越大，对文本类别分布的影响也越大，因此选择 KL 最大的 K 个词作为特征。期望交叉熵的定义如下：

$$ECE(t) = p(t) \sum_{i=1}^{|c|} p(c_i|t) \log \frac{p(c_i|t)}{p(c_i)}$$

这里 p(t)表示在训练数据中含有特征 t 的文本的概率，$p(c_i)$为训练文本中类别c_i的概率，$p(c_i|t)$表示当特征 t 出现时文本属于c_i的概率。

卡方检验的基本思想是观察实际值与理论值的偏差来确定理论正确与否。如果原假设为两个变量相互独立，然后观察实际值与理论值的偏差，若偏差足够小则接受原假设，即认为两个变量确实是相互独立的；如果偏差大到一定的程度，不可能是偶然发生或者仪器测量不

准确所导致的，那么就否定原假设，即认为两变量不是相互独立的而是相关的。

在特征选择的过程中，一个很合理的假设是提取在一个类别中出现概率高的特征 t 并且特征 t 在另一个类中出现的概率低；并且为了避免选择算法偏向于选择低频词，还需要考虑特征 t 的出现概率的绝对大小。考虑到这两个方面，本文作者设计了一个特征选择算法 ADMMR，对于每个特征 t：

（1） 找出其在类中出现的最大概率$p(t|c_{max})$，在类中出现的最小概率$p(t|c_{min})$

（2） 计算特征 t 的$ADMMR(t) = [p(t|c_{max}) - p(t|c_{min})] \log \frac{p(t|c_{max})}{p(t|c_{min})}$

（3） 选择出 ADMMR 最大的 K 个特征；并在各个类之间特征的平衡。

此方法避免了信息增益方法中考虑特征未出现所带来的负面效应，也避免了互信息方法考虑特征与每个类的互信息关系而对特征的优势值所带来的削弱作用。在 ADMMR 中，$p(t|c_{max})$能够提升其所在类c_{max}的分类准确度，而$p(t|c_{min})$能够提升其所在类c_{min}的分类的召回率。如果类的集合比较多，则可以在类之间做一个特征的平衡，不过不做特征平衡分类的效果也是相当不错的。

同时因为简繁文本转换的特殊性，那些靠近待转换字的特征应该赋予一个相对较高的优先值，远离待转换字的特征可以赋予相对较低的优先值，在此思想下可进一步对特征选择算法进行优化。简繁转换特征的另一个特点是对于包含一对多简体字的无歧义词组可以根据规则直接转换，因此特征中不需要含有这些无歧义词组；同时若包含这些无歧义词组特征则会减弱其他特征的作用；因此为了避免这一问题作者将无歧义的词组数据进行了过滤而不作为特征。

4 实验

本文的主要思想是使用规则加组合统计模型来解决简繁转换中的难点问题：一对多简体字转繁体字。本文所使用的统计模型有：SVM 模型，最大熵模型以及贝叶斯模型，而为了验证组合模型解决这一问题的有效性需先验证 SVM 模型，最大熵模型以及贝叶斯模型都可以在一对多简繁转换问题上取得一个不错的效果。

本文不只是将这三种模型应用于简繁转换问题，而是对各模型的训练过程进行了一定的改进：提出了 ADMMR 特征选择方法；提出将 tf-idf 应用于最大熵模型，并且对 SVM 模型和最大熵模型均采用了 3 种不同的特征选择方法，最终形成了 3 个 SVM 模型文件和 3 个最大熵模型文件。根据这一逻辑顺序，本章将会先进行 ADMMR 特征选择方法的有效性验证以及 tf-idf 应用于最大熵模型的有效性验证实验；然后再进行组合模型对简繁一对多转换问题的有效性验证实验。

4.1 ADMMR 特征选择实验

为了体现方法的有效性，本实验所使用的数据为公认数据：搜狗实验室所整理的搜狗新闻文本分类数据。搜狗新闻数据一共包括汽车，财经，IT，健康，体育，旅游，教育，招聘，文化，军事等十类，因一对多简体字所对应的繁体字一般不超过四个，所以在测试 ADMMR 的有效性时只选择三或四类数据进行测试。在本实验中选择了财经、健康以及军事这 3 个较为平衡的类别进行测试，其中每个类别训练数据为 1600 个文本，待测试数据为 390 个文本；评价指标采用分类准确率来比较。

测试的特征选择方法为：IG 方法（信息增益法），互信息法，卡方校验法，期望交叉熵，ADMMR 以及无特征选择法；特征选择数量为 8500、6000 和 3500；使用的分类模型为 SVM 模型。实验结果如下：

特征数	类别	IG	MI	ECE	X2	ADMMR	不选择
3500	财经	0.2385	0.9256	0.9615	0.9487	0.9462	0.9231

	健康	0.9872	0.5103	0.9692	0.9718	0.9769	0.9718
	军事	0.959	0.4744	0.9821	0.9821	0.9872	0.9821
6000	财经	0.3128	0.9462	0.9436	0.941	0.9359	0.9231
	健康	0.9872	0.641	0.9667	0.9744	0.9744	0.9718
	军事	0.9692	0.5949	0.9872	0.9872	0.9897	0.9821
8500	财经	0.3103	0.941	0.9282	0.9333	0.9436	0.9231
	健康	0.9846	0.6744	0.9641	0.9641	0.9667	0.9718
	军事	0.9564	0.6615	0.9897	0.9872	0.9846	0.9821
平均准确度		0.745	0.7077	0.9658	0.9655	0.9672	0.959

表 1 各特征选择方法准确度对比 1

通过本实验发现 IG 和 MI 特征选择算法的分类准确度都不是很好，IG 的准确度为 74.5%，信息增益的分类准确度为 70.77%。这两种方法最不理想的地方是其对部分类的分类准确度太低，导致这种现象出现的原因是 IG 和 MI 都没有选择出最能代表类别的特征且未能在类别之间有个较好的平衡。ECE，X2 和 ADMMR 的分类准确度都为 96%这一个层级，超过了将所有词组作为特征（共 43777 个特征）所能达到的分类准确度。而在 3500,6000,8500 这 3 个特征数层级上，各方法的分类准确度都差不多，因此对于这个语料的分类问题，完全可以使用 3500 维的特征来代表文本；并且当特征选择方法采用 ECE，X2 和 ADMMR 时都能达到将所有词都作为特征时的分类准确度。

同时为避免因简繁转换文本的特殊性而导致特征选择方法在一对多简繁转换问题不具备上述特征，特选择一对多简体字“干”也作为特征选择试验的效果比较，其中将全部词作为特征时特征维数为 91869。

特征数	类别	IG	MI	ECE	X2	ADMMR	不选择
3500	干	0.8546	0.7435	0.9121	0.921	0.92	0.9031
	幹	0.8736	0.7503	0.9292	0.9218	0.921	0.945
	乾	0.8835	0.6744	0.9243	0.913	0.9231	0.8987
6000	干	0.8673	0.7462	0.931	0.882	0.9359	0.9031
	幹	0.8912	0.7055	0.9123	0.9344	0.9144	0.945
	乾	0.8543	0.7032	0.9272	0.921	0.9097	0.8987
8500	干	0.8303	0.8125	0.9182	0.9333	0.9332	0.9031
	幹	0.9025	0.6378	0.935	0.915	0.9187	0.945
	乾	0.8637	0.6725	0.8945	0.9262	0.905	0.8987
宏平均		0.869	0.7162	0.9204	0.9186	0.9201	0.9156

表 2 简繁转换问题各特征选择方法准确度对比

从上述实验可以看出，ADMMR 和卡方检验，期望交叉熵在特征选择上具有最好的效果，因此在综合模型前的特征选择将采用这三种方法。对于每种方法所表示出来的训练文本再采用 SVM 和最大熵模型来学习，从而产生分类模型。

4.2 最大熵模型之 tf-idf

最大熵模型最开始对每个特征使用的值为 0 或者 1，但是在文本分类领域显然每个特征的 0-1 值所能代表的信息不够丰富，因此作者在实验过程中就考虑将每个特征的值采用 tf-idf，并且实验表明采用 tf-idf 后，分类准确度有所提升。

本次实验仍然采用前一节所使用的实验数据，而根据前一节的实验结果发现特征维数为

3500 时基本能代表整个文本，因此本节的实验特征维数固定为 3500 维，特征选择方法则采用 ADMMR。另外 0-1 值如此规定：当文本中出现了该特征时则相应的特征值设置为 1。本实验中迭代次数设置为 100，在实验时指明参数为“-real”表示特征值为实数。第一个实验仍然是针对搜狗新闻语料并且也是选取财经，健康和军事这三类。

特征数	类别	tf-idf	0-1 值
3500	财经	0.941	0.9205
	健康	0.9744	0.9744
	军事	0.9923	0.9872
平均准确度		0.9692	0.9607

表 3 tf-idf 与 0-1 值特征准确度对比 1

从该实验结果可以看出最大熵分类模型采用 tf-idf 作为特征值时分类准确度提高了约 1%，而且训练所耗费的时间基本一样。这一结果也证明了作者的想法是正确的，在做文本分类时 tf-idf 可以比 0-1 值蕴含更多信息。而比较前一节分类所采用的 SVM 分类模型，可以得知最大熵分类器和 SVM 分类器的准确度不相上下。由于这两个分类器是不同的原理，因此作者在实验的过程中就想到如果将这两种分类效果不错的分类器组合起来，是否能够弥补对方的缺点？这个组合模型的实验将在下一节进行叙述。但是为了进一步证明 tf-idf 的有效性，作者又将 tf-idf 应用于一对多简体字“制”的分类数据。实验结果如下：

特征数	类别	tf-idf	0-1 值
3500	制	0.9634	0.9405
	製	0.9096	0.8832
宏平均		0.9365	0.9119

表 4 tf-idf 与 0-1 值特征准确度对比 2

实验结果显示，最大熵分类模型使用 tf-idf 作为特征值比使用 0-1 值作为特征值分类结果更有效。

4.3 SVM、MaxEnt 以及 Bayes 组合模型

经过前两节的实验证明了 ADMMR 特征选择方法的有效性和 tf-idf 特征值对最大熵分类模型的有效性，因此本节将叙述组合模型的相关实验。本实验中特征选择方法采用 ADMMR，ECE 以及 X2 这三种特征选择方法，而对于每个特征选择方法所选择表示出来的训练文本采用 SVM 模型和最大熵模型进行学习，因此就会产生 3×2＝6 个分类模型。对于这 6 个分类模型的结果采用投票的模式来组合，这就可能会产生平局的情况。对于平局的情况如果采用随机方法去选择则只有最高不超过 50%的正确率，作者因此引进了训练和分类都比较简单并且分类效果的准确率远大于 50%的 Bayes 模型。

Bayes 分类模型的分类原理又不同于 SVM 和最大熵，因此在出现平局时恰好可以作为最终的裁判者。因此这里我们需要知道 Bayes 模型在文本分类中的准确度和产生平局时 Bayes 模型所带来的准确度提升有多少。第一个实验使用搜狗的新闻语料数据中的财经，健康，招聘,军事和教育 5 个类别的数据，本次实验结果如下：

特征数	类别	SVM	MaxEnt	Bayes	组合模型
3500	财经	0.9	0.8974	0.8462	0.9103
	健康	0.8821	0.8974	0.7103	0.9103
	教育	0.9	0.8949	0.8179	0.9154
	招聘	0.8205	0.8564	0.9	0.8667

	军事	0.9769	0.9692	0.8718	0.9769
平均准确度		0.8959	0.9031	0.8292	0.9159

表 5 分类模型准确度对比

从这 5 类的分类结果可以看出，采用组合模型的分类准确度比 SVM 高出 2%，比最大熵高出 1.2%，比 Bayes 模型高出了将近 10%。同时我们也可以看到当类别数为 5 时，SVM 模型在对招聘类的数据进行分类时出现了 82%的相对不理想的准确度，并且根据波动指数：$\mathrm{Wave(f)} = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - E(p_i)\right)^2$，可知：Wave(SVM)=0.00249，Wave(maxent)=0.001336，Wave(Bayes)=0.004276，Wave(组合模型)=0.001241，可知组合模型的稳定性最好。且结合作者另做的实验发现组合模型在类别数越多时，分类稳定性更好，相对优势更明显。

4.4 规则加组合统计模型实验

简繁文本的双向翻译问题中的繁体转简体只需要注意两点：是否术语转换和繁体转简体时特别词语的转换。中科院的分词系统 ICTCLAS[24]支持繁体分词，因此对繁体文本进行分词之后，再根据用户要求是否进行术语转换，然后再判断该词是否需要特殊转换，最后为一一转换，此部分的功能准确度能够接近 100%.

简繁文本双向翻译的重点难点还是在一对多简体字转换为繁体字的问题，本文提出的规则加组合统计模型的方法，即：收集一定可成词简体字词组与对应繁体字词组的词库；另外对于不能成词的一对多简体字采用 SVM，最大熵以及贝叶斯模型的统计模型学习词的上下文；在词库无法解决时，采用统计模型来处理未能成词的数据。在准备组合模型的训练数据时，需先剔除可成词的部分以免产生干扰。为此作者准备了无成词的训练数据，本实验时针对“干”字，其中“干”，“幹”，“乾”这三类训练数据都为 1500 个文本，测试数据分别为 500、400 和 500，另一类“榦”数量太少不参与分类，测试数据实验结果如下：

特征数	类别	SVM	Maxent	Bayes	组合模型
3500	干	0.7735	0.7856	0.6035	0.821
	幹	0.7922	0.7903	0.7824	0.7955
	乾	0.7643	0.7569	0.5306	0.8025
宏平均		0.7767	0.7776	0.6388	0.8063

表 6 不成词情况分类模型性能比较

实验结果显示，组合模型的分类准确度和稳定度都是最好的，显示组合模型在抽取上下文以解决一对多简繁转换问题也有一定的优势。

同时为了测试词库加组合模型在解决简繁转换一对多问题上的优势，作者收集了台湾的新闻数据及古龙的繁体小说进行简繁转换测试。所选取的相比较系统为谷歌翻译[25]、同文堂[26]，快典网[27]以及厦门大学的简繁汉字转换系统[28]。鉴于简繁转换的工作量之大以及互联网数据的准确性需要人工验证，而同时为了获得最准确的转换效果，本实验不可能针对所有一对多简繁字进行实验。因此，本实验只针对“干”字进行实验，共有测试文本 14034 个，其中可成词数据约为 13300 左右，其他数据中的“干”字未见明显成词。作者收集到了“干”字的相关短语 650 组，组合模型部分的训练数据则为之前的训练数据和测试数据的合并集，即“干”，“幹”，“乾”这三类训练数据分别为 2000、1900 和 2000。实验结果如下：

	谷歌翻译	同文堂	快典网	厦门大学系统	本系统
转换正确数	12893	12588	11087	13687	13836
未成词转换正确数	415	199	187	555	609

总正确率	0.9187	0.897	0.79	0.9753	0.9859
未成词正确率	0.5654	0.2711	0.2548	0.7561	0.8297

表 7 一对多转换综合比较

厦门大学的系统相对来说是一个比较好的系统，但是经过分析结果发现该系统在未能成词时转换“干”字的问题还是比较大的，特别是在转换表示做事意思时的“幹”，表示结拜而来的亲属“乾”以及插手意思的“干”字时错误较多；另外厦门大学的系统其词库存在问题，或者是未收录或者是收录错误，例如其系统存在“搞單乾”，“提乾”，“實乾”，“根干”等转换错误。为体现本系统的作用，特别例举出如下转换成功的例子：

部分语言环境	目标繁体字	实际繁体字
酸雨正式的名称是为酸性沉降，它可分为“湿沉降”与“干沉降”两大类	乾	乾
倘能把附在食物表面的微生物水份抽干，微生物就难以生存	乾	乾
中国最早有饴、饧、糖等字，都是以糯米为原料，稀的叫饴，干的叫饧、糖。	乾	乾
再装入小沙锅焖煮（每锅一般二至三两），水快干时放入香肠	乾	乾
当胶水干了之后，便可切割成独立的铅笔	乾	乾
播种时需保持泥土湿润，不要积水也不要过干	乾	乾
加以自然环境恶劣，不是太干就是过湿	乾	乾
使之打断水分子之间的氢键，脱离原本附着的表面(即蒸发)，头发因而变干	乾	乾
将水煮干以形成盐结晶	乾	乾
凡是遇到机会就不要丢，就是要坚持，要干起来，要体现改革开放	幹	幹
你们对我们干了‘不该’的事	幹	幹
像我们常见的在一些经常干重体力拉动	幹	幹
1943 年他总结推广“拥干爱兵”经验	幹	幹

表 8 未成词时本系统转换成功示例

从表 8 可以看出，本系统在这些例子上使用上下文知识对待转换简体字“干”进行了正确的转换。本系统的转换错误主要集中于部分数据计算机无法从上下文判断出来，例如对于句子“我的脑子里有些不好的想法，而且感觉到他要干票大的”。对于这个句子中间的“干”字，周围的词对判断目标繁体字基本没有任何帮助；另外一个就是测试数据中有不少地名，人名中间含有“干”字，这种除非加入字典否则很难解决，但字典也不可能收集所有的这些词。例如测试数据中有“松赞干布”、“布拉干萨王朝”、“德干高原”、“东干族”、“乌干达”等人名地名或者固定用法，这些词相对来说是比较难处理的。上述两个问题是简繁转换系统要趋于完美所难以克服的问题！

5 总结与展望

随着两岸交流越来越频繁，大陆地区与台湾地区使用的汉字差异给交流带来了障碍，为了消除这种障碍，两岸甚至全世界都在研究如何将简繁文本进行互相翻译。简繁文本进行互相翻译的重点难点问题是一对多简体字转繁体字的问题，因为存在这种一对多关系，因此需要根据上下文来决定简体字所要转换到的目标繁体字。已经有不少机构在研究这一问题，但是转换的效果都不是很好，本文提出了使用组合统计模型结合词库的方法来解决这一问题。

一个简体字可以转换为多个繁体字，而转换的过程中需要根据上下文来判断到底是转换到哪一个繁体字，因此可以看出这是一个分类问题。而 SVM，最大熵都在分类问题上取得了不错的效果，同时贝叶斯作为一个简单的分类器也可以产生意想不到的好效果。因此，作者提出将这三个模型组合起来以弥补各自的优势，实验表明组合分类器确实能够有很到的分类效果和分类稳定性。作者同时提出了一种新的特征选择算法 ADMMR，实验表明这种特征选择算法对分类问题有不错的效果，并且在简繁一对多转换问题上同样有不错的效果；此外，作者还同时提出了使用 tf-idf 作为最大熵分类器的特征值。实验表明在文本分类问题上使用 tf-idf 比使用 0-1 值更能获得好的分类效果。

规则加组合模型的方法虽然能很好地解决简繁转换问题，但是就如同作者在文中所写，分类模型需要优质的训练数据，而现在因为两岸未在简繁对应关系上做出明确规定，因此训练数据总是存在错误并且从网络所获取的数据也是良莠不齐，导致了本文中作者只能选择性地针对部分数据做实验；另外简繁转换问题始终有难以克服的问题，例如上下文不包含有利于判断目标繁体字的内容，或者该简体字存在于人名、地名或者其他固定用法中，因此这些始终都是很难完全克服的障碍，只有随着两岸互动增加对相关问题作出规定才能有效解决剩下的问题。

参考文献：

[1] 付永和．汉字简化五十年回顾[J]．中国语文，2005(6)：537-539．

[2] 李义琳．简繁汉字中的“一对多”[J]．寻根，2009(05)：14-17．

[3] 汤吟菲.《简化字总表》繁简字对应关系的注释说明[J]．郧阳师范高等专科学校学报，2010(04)：33-35．

[4] 魏励．简化字与繁体字的几点思考．商务印书馆，2010(04)：33-35．

[5] 龙城顺.有的“於”不能简化作“于”[J].语文建设,2001,(04).

[6] 苏培成.“发”字的尴尬[J].语文建设,2001,(12)..

[7] 王宁，王晓明.两岸四地汉字的转换与沟通[A].第三届两岸四地中文合作论坛[C].2005.

[8] 王宁.基于简繁汉字转换的平行词语库建设原则[A].第四届两岸四地中文数字化合作论坛[C].2007.

[9] 王晓明，魏林梅.“谈简繁转换的几个关键问题”,5TH CDF 研讨会数位社群双效(CD2E),2008年12月24.

[10] 郑国政.基于现有软件进行中文简繁体转换的方法[J].电脑知识与技术,2007,(03).

[11] 冯霞.中文繁简转换及其转换工具[J].计算机教育,2007,(05).

[12] 王立军,王晓明.简繁对应关系与简繁转换[J].中文信息学报,2013,(07).

[13] 刘汇丹,吴健. 基于词语消歧的分层次汉字简繁转换系统[J].中国语言战略,2012,1(1)25-35.

[14]辛春生，孙玉芳. 简繁汉字转换系统的设计与实现[J]．软件学报，2000,(11)．

[15]辛春生，孙玉芳.汉语简繁体转换与语词切分[J].小型微型计算机，2000,21(9):982-985

[16] Slavam. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer ”, IEEE Transactions on ACOUSTICS, SPEECH, and SIGNAL PROCESSING, VOL. ASSP-35, NO. 3, MARCH 1987, pp 400-401.

[17] 李民祥，吴世弘，曾议庆，等．基于对照表以及语言模型之简繁字体转换 [J]. 中文计算语言学期刊，2010,15(1):19-36

[18] Fai Wong, Mingchui Dong. Chinese Conversion Based on Statistic Model，5TH CDF 研讨会数位社群双效(CD2E),2008年12月24.

[19] Aoe,J.An Efficient Digital Search Algorithm by Using a Double—Array Structure[J]IEEE Transactions on Software Engineering.1989,(09).

[20] Tianyong Hao,Chunshen Zhu.Simplified-traditional Chinese character conversion based on multi-data resources:Towards afused conversion algorithm[C]//Preceedings of the 2nd International Conference on Next

Generation Information Technology(ICNIT).2011:50-56.

[21] YidongChen,XiaodongShi,ChangleZhou.A Simplified-Traditional Chinese Character Conversion Model Based on Log-Linear Models[C]//Proceedings of 2011 International Conference on Asian Language Processing(IALP).2011:3-6

[22] Xiaodong Shi,Xiuping Huang. Key Problems in Conversion from Simplified to Traditional ChineseCharacters，International Conference on Asian Language Processing,2011.

[23] Martin Hepp, Katharina Siorpaes, Daniel Bachlechner, "Harvesting Wiki Consensus: Using Wikipedia Entries as Vocabulary for Knowledge Management," IEEE Internet Computing, vol. 11, no. 5, pp. 54-65, Sep./Oct. 2007.

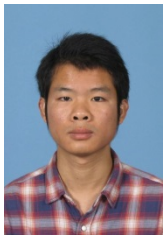
[24] 中科院计算所ICTCLA2009, http://ictclas.org/index.html

[25] 谷歌翻译. http://translate.google.com.hk/#zh-CN/en/.

[26] 同文堂.http://tongwen.openfoundry.org/.

[27] 快典网. http://ft.kdd.cc/.

[28] 厦门大学简繁汉字智能转换系统. http://jf.cloudtranslation.cc/s2t.html.



庞祯军（1987—），男，硕士，主要研究方向为意见抽取，信息抽取，自然语言处理，pzj_636484@163.com

姚天昉（1957—），男，博士，副教授，硕导，主要研究方向为意见挖掘、信息抽取、机器学习、自然语言处理等，

yao-tf@cs.sjtu.edu.cn