一种基于认知情景框架的文本分类方法*

李月伦[1] 李湘[2] 常宝宝[1] 袁毓林[2]

[1]北京大学 计算语言学研究所，北京，100871

[2]北京大学 中文系，北京，100871

教育部计算语言学重点实验室

Email：lyldtc.student@sina.com, lixiang2610@126.com, chbb@pku.edu.cn, yuanyl@pku.edu.cn

摘要：在文本分类领域，常用的特征选择方法（如文档频率）是基于概率统计信息的。本文从一个全新的角度，即基于认知情境的语义框架的角度进行特征选择，这种方法可以较准确地抓住文本类别的特征，对提高分类精度与分类效率起到促进作用。由于基于机器学习的文本分类方法近年来得到了广泛的关注与快速的发展，本文运用最大熵分类器，以“罢餐”类文本为例进行文本分类实验，召回率可达 96.8%。

关键词：认知情境 语义框架 最大熵 文本分类

A Text Categorization Method Based on Cognitive Situations

Li Yuelun, Li Xiang, Chang Baobao, Yuan Yulin

Institute of Computational Linguistics, Peking University, Beijing, 100871

Abstract: In the field of Text Categorization, the common used feature selection methods such as the document frequency are based on probability and statistics. In this paper, from a brand new perspective, we select features base on cognitive and semantic framework. This method can grasp the characteristics of the classification accurately, which can improve the classification accuracy and efficiency. Because the classification method using machine learning have gained widely attention and usage, we use maximum entropy classifier and take the “strike” texts as example to do some text categorization experiments, which the recall can achieve 96.8%.

Keyword: cognitive situations, semantic frame, maximum entropy, text categorization

1 引言

文本分类(Text Categorization)是指根据文本的内容，通过某种自动分类算法，由计算机把文本判定为预先定义好的类别[1]。

中文文本分类方法可分为如下两类[2]:

(1) 基于外延的分类方法。该方法不关心文本的语义，仅根据文本的外在特征以及机器学习理论对文本进行分类，例如各种统计方法。

(2) 基于语义的分类方法。该方法根据全部或部分文本的语义进行分类，但此方法的发展受到了自然语言处理技术的制约。

用统计方法对中文文本进行分类，首先需要将文本进行预处理（包括去除无用标记，进行分词等），并去掉噪音和词频很高或者很低的垃圾数据，然后统计词频，将文档表示成一个以特征项的权重为分量的向量表示$(w_1, w_2, w_3, \ldots, w_n)$，其中$w_i$为第$i$个特征项的权重，特征项通常

* 本文工作受到自然科学基金(60975054)、社会科学基金(06BYY048)的支持，特此致谢。

由中文词构成，权重 w_i 一般用词频表示。

过高维数的特征向量以及海量的训练样本是文本分类的两大特点[3]。由于存在数量巨大的中文词，因而表示每个文档的特征向量的维度都是巨大的，因此需要对文档的特征向量进行降维。降维的一个主要技术是特征选择。

在文本分类系统中，特征选择是指去除信息量较小的词，以提高分类效率和降低计算复杂度[2]。特征选择可以保留有区分能力的特征，去除冗余特征，达到减小文本的特征向量维数的目的。同时，具有区分能力的特征可以有效提高分类系统的效率和精度。因此，特征选择在文本分类系统中是至关重要的[1]。

常用的特征选择方法有：文档频率、信息增益、互信息、χ^2统计量、期望交叉熵、文本证据权和几率比等。这些方法的共同特点是利用概率统计的信息来判断某一个特征的重要程度。从已有研究的效果来看，此类方法确实帮助文本分类系统取得了较高的分类精度，但也应该看到，统计的策略在实际应用当中已经得到了几近极致的发挥，我们希望能在追求技术层面改进的同时，在视角转换和思路调整上下功夫，突破分类精度的瓶颈。

出于以上的考虑，我们希望能通过语言本体知识的介入，从文本的语义内容出发寻找可以代表文本类别的关键特征。具体而言，本文将引入"基于认知情境的语义框架"这一概念，从情境角色、情境网络组成等方面对文本的语义内容进行具有区别性的描写和概括，并在此基础上更加准确地抽取出能够代表文本所属类别的关键词，从而构成相应的文档的特征向量表示。

在算法和程序实现上，本文将采用最大熵模型作为分类器。基于机器学习的文本分类技术具有降维等特点，因而在近年来得到了很大的发展[4]。其中，最大熵模型在已知部分信息的条件下，产生符合该已知信息的最不确定的或最随机的预测分布，使未知分布尽可能地均匀分布，李荣陆等证明了基于最大熵模型进行中文文本分类可以去的较好的效果[5]。

为展示本文所提方案的可行性，我们将选取与"罢餐"事件有关的文本进行文本分类过程的全程演示。其中预先设定的目标是，比较精确地识别出网络中出现的真实的、动态的、带有煽动"罢餐"倾向的文本。由于我们希望尽可能正确地识别出"罢餐"类文本，因此，在实验过程中，召回率是首要考虑的目标。

2 基于认知情境的"罢餐"语义框架描述

2.1 认知情境和语义框架

依据Winograd (1983)[6]的见解：语言使用是一种以知识为基础的交际过程，人说出或理解一句话时，在大脑中有一个关于所描述的外部世界中的事物或事件的心理映象，可以称之为内部语言；而人处理语言的过程就是把外部语言转化为内部语言，经过加工后再由内部语言转化为外部语言的过程。以这一观念为理论背景，袁毓林等（2005，2008，2009）[7][8][9]提出了一种认知语言学研究的计算范式与技术路线，其中的先行步骤就是针对自然语言进行认知建模（cognitive modeling）——对有关语言现象所指谓的事件、关系或状态作出认知假设，对它们所涉及的语义情境（semantic situations）进行分析和描写，进而在语义概念层面上建立起对有关事件、关系和状态的模型化概括。这种模型化的概括就是所谓的"认知情境"。（多个相关的"认知情境"可以构成一个"情境网络"。）

概念层面的事件、关系和状态在语言层面上通常对应为具体的谓词（动词或形容词），因此，认知情境和情境网络总是可以兑现为（一个或一组）谓词及其论元之间的组配模式。这种组配模式可以称之为"认知情景框架"，或者简称"语义框架"。袁毓林（2008）[8]从计算机处理的需要出发，详细地研究了汉语动词论元结构的论元属性、论旨属性、语法特征、语义特征以及配位方式，并依据自立性、使动性、感知性、述谓性、变化性、受动性、渐成性、关涉性、类属性等动态语义特点，把汉语动词的论元分为施事、感事、致事、主事、受事、与事、结果、对象、系事、工具、材料、方式、场所、源点、终点、范围、经事、原因、目的、时间、路径、话题、说明等语义角色。袁文的研究正是本文为特定认知情境编写语义框架的基础。

2.2 "罢餐"情境的语义框架描述

下面以"罢餐"情境为例对语义框架的具体构建进行说明。"罢餐"一词可以在人脑当中激活一个罢餐事件的心理映像，将其模型化以后即得到了"罢餐"的认知情境。在这个情境中，特定的事件参与者以特定的关系相互依存。依据参与者之间的关系为各个参与者分派语义角色，我们就能得到"罢餐"事件的语义框架，如下所示：

<情境名称>：罢餐

<情境定义>：

就餐者（通常为学生或员工）为实现某种要求或表示抗议而集体拒绝在特定的膳食供应点就餐。

<情境角色>：

施事 A：某饮食供应处的常规就餐者；

与事 D：涉及餐饮管理的职能部门；

受事 P：遭受罢餐抵制的餐饮供应者；

场所 L：饮食供应点，通常和 P 同形；

方式 M：拒绝就餐的集体性抵制方式；

原因 RN：导致罢餐的原因；

目的 AI：实现某种要求或表示抗议；

<情境网络组成>

前提事件：发起罢餐前的组织和准备；

后续事件：罢餐发生后所造成的影响；

注：其中的<情境网络组成>显示了"罢餐"这一基础事件与其他相关事件之间的逻辑联系，这一项目可以根据文本分类的直接目的和精度要求进行删减或扩充。

3 最大熵模型

最大熵模型可以用于解决分类问题。首先，需要建立条件概率模型 $p(y \mid x)$ 进行统计分类。其中，x 代表已知信息，y 代表该问题的分类结果。$H(p)$ 代表分布 p 的条件熵，C 表示满足条件的所有概率分布。那么，满足已知条件并且熵值最大的分布如(3.1)所示

$$\begin{cases} p^{*} = \arg\max_{p \in C} H(p) \\ p\text{需要满足的约束条件} \end{cases} \tag{3.1}$$

式(3.1)中提到的约束条件即是特征，特征函数 $f_{x',y'}(x,y)$ 的形式如(3.2)所示

$$f_{x',y'}(x,y) = \begin{cases} 1, \text{如果} y = y' \text{且} x = x' \\ 0, \text{其他} \end{cases} \tag{3.2}$$

如果已知信息满足某种条件，特征函数为 1，满足该条件下的所有信息都会在统计建模时给予考虑。如果已知信息不满足某个条件，特征函数为 0，该条件下的所有信息都不会予与考虑。

比如，以基于字符标注的汉语分词方法提取出的特征为例，存在如下(3.3)所示的特征函数

$$f(x,y) = \begin{cases} 1, \text{如果} t_{i-1} = LL \text{且} t_i = RR \\ 0, \text{其他} \end{cases} \tag{3.3}$$

如果 C_{i-1} 被标记为 LL 且 C_i 被标记为 RR，那么这个特征将会对预测的概率分布做出贡献。

这样，式(3.1)即可转化为在某个约束条件下的最优化问题[10]（optimization problem），通过拉格朗日条件约束（lagrangian constraint）求极值推导，可以得到最终所求的条件概率分布如(3.4)所示

$$p(y \mid x) = \frac{1}{Z(x)} e^{\sum_{i=1}^{k} \lambda_i f_i(x,y)} \tag{3.4}$$

其中，λ_i 是特征函数 $f_i(x,y)$ 的权重，用来表示该特征的重要性程度，需要通过训练得出。

k 是特征的个数。$Z(x)$ 通常被称为分割函数，是一个标准化因素，用来确保 $\sum_{y} p(y \mid x) = 1$。

指数分布函数是凸函数，所以最大熵模型的参数会具有全局最优解。全局最优解可以通过梯度下降等最优化算法求得。

因为指数分布本身是一个线性分布，每个特征也表示线性函数的一维。特征越多，对客观数据的模拟将会越接近，模型的精度也将越高，但与此同时，函数的空间维数也将不断增长，求得模型参数的训练时间也会随之增加。

4 文本特征的设置

利用机器学习的方法进行文本分类的首要任务和关键问题是特征选择和特征抽取[10]。

根据上文中所描述的认知情境框架，我们从网络上收集了大量包含有“罢餐”一词的文本，并人工对每篇文本是否属于“罢餐”类文本进行了分类。分类的依据为：鼓励他人进行罢餐运动的煽动性类文本以及对“罢餐”运动给予报导或评论的新闻性文本属于“罢餐”类文本，其他文章属于

“非罢餐”类文本。

4.1 基于语义框架的特征词提取

从上文描述的基于“罢餐”情境网络的语义框架中，我们可以提取出“罢餐”类文本的一些特征，比如策划、煽动等（情景网络组成中的前提事件）。

作为对语义框架的改进，我们引入主题谓词 V 和情境角色 A、D、P、M、RN、CM 作为基本事件匹配过程中的操作单位。

主题谓词 V：罢餐，拒绝去食堂，拒绝到食堂、不再去食堂、不再到食堂，不吃食堂、别去食堂、不要去食堂、不要在食堂……
施事 A：学生、同学、师生、员工、职工、职员、学校、学院
受事 P：食堂、饭堂、餐厅、饮食部、餐饮部、负责人，承包
与事 D：学校、学院、校方、领导、后勤处、后勤部、膳食部
场所 L：食堂、饭堂、餐厅、餐饮部、餐饮中心
方式 M：集体、抗议、静坐、示威、游行
原因 RN：饭菜、伙食、质量、差、菜式、分量、少、不够，吃不饱、肉少、不合算、价格、高、贵、脏、不干净、卫生、堪忧、不合格、吃出
伴随 CM：自备、方便面、泡面、叫外卖、订餐、食堂冷冷清清、空空荡荡

我们认为，引入情境角色作为基本事件匹配过程中的操作单位有两个好处：一是，语言项目相同类型的事件，必然对应相同情境模式，而情境角色的数量大大小于单个特征词的数量，因此，以情境角色为操作单位来进行字符串匹配，效率更高。二是，这样的处理还能避免同一情境角色以不同语言表达形式多次出现从而导致权重的虚高。我们假定当前输入了这样一个文本：“地板很脏、餐具不干净，还有同学从菜里吃出了苍蝇。……卫生状况堪忧，上次检查就不合格。”尽管这段话也许与“罢餐”毫无关系。但是其中出现了多个“罢餐”事件的特征词，如果以特征词为匹配操作的单位，那么，这些特征词的权重累加，很可能使得文本间的相似度的计算出现偏差，误把输入文本当作“罢餐”文本。

4.2 预选特征的提纯性过滤

根据上文所述的关于“罢餐”类文本的认知情境的语义框架以及相应的改进措施，我们可以得到很多“罢餐”类文本应有的特征，但将上文中提到的特征全部纳入考虑范围，不仅表示文档的特征向量维数会较高，而且特征向量会非常稀疏，不会对文本分类起到促进作用。因此，基于从语义框架中提取出的特征，我们对“罢餐”类文本进行了相应的分析，得到了一些该类文本具有的普遍特征，舍弃了只有在个别“罢餐”类文本中才会出现的特征。

从分析中，我们不难得到“罢餐”类文本中经常出现的一些词语，比如“难吃”，“贵”，“忍不了”等（原因 RN）。此外，一些煽动性较强的发布在学校 BBS 上的“罢餐”类文本，会较多出现如“团结”，“组织”，“帖子”等词（方式 M）。除了煽动性“罢餐”类文本外，含有“罢餐”一词的新闻报道类文本也在总文本中占有不少比例。该类文本针对“罢餐”事件进行报导，多会出现如“方便面”（伴随 CM），“活动”（前提事件）等词。

通过数次实验的验证，得到了如下可使分类精度达到较高的关键词特征：

罢餐 罢饭 团结 饭菜 差 少 不够 高 贵 涨 降 方便面 泡面 改善 活动 抗议 难吃 不满 食堂 饭堂 餐厅 量 抵制 拒绝 同学 短信 帖子 下调 过分 游行 忍不了 不忍 组织 活动 不去

4.3 词频信息因素的处理

在本文中，词频信息为某一个特征词在一篇文本中出现的次数。

对于某些“罢餐”类文本，只出现上述特征词中极少的一部分，而出现的特征词在该文本中出现了若干次，比如“方便面”一词：“他们的早餐以方便面和饼干为主”，“早上吃方便面充饥”。为了凸显该关键词对于文本分类的重要程度，因而在提取特征时考虑了词频信息。

此外，对于较普遍出现在所有文本中的特征词，比如“罢餐”一词，设置词频信息也可以强调该特征词的词频对“罢餐”类文本的贡献。由于每篇文本中都包含“罢餐”一词，因此文本中是否包含“罢餐”一词，理论上已不能对文本的分类产生较大影响，对于煽动性较高的文本，一般会出现多次的“罢餐”，而在一些“非罢餐”类文本中，比如网友的博文中，可能会提到自己参加过的一次“罢餐”运动，这种情况下，“罢餐”一词一般只会出现 1～2 次，与煽动性“罢餐”类文本形成了鲜明对比。因此，特征词的频数可作为文本是否为“罢餐”类的重要衡量。

考虑到如果用特征词出现的次数作为特征时，由于文本数量比较大，且对于每个特征词，各篇文本出现该特征词的次数不定，因此会造成特征数量剧增且每个特征出现次数过少导致的特征稀疏问题，因此，将词频信息采用分段的方式进行处理。具体的处理方式为：特征词不出现在某篇文本中，特征设为 0；特征词在某篇文本中出现 1～5 次为低频，特征设为 1；出现 6～10 次为中频，特征设为 2；其余情况为高频，设为 3。

4.4 其他限制条件的设置

通过观察分析收集到的文本，可以发现“罢餐”类文本，尤其是带有煽动性的“罢餐”类文本，多在文本的标题处就出现“罢餐”一词。此外，一些呼吁罢餐的文本，在题目中写明“罢餐”后，在正文中开始描述食堂的种种缺点，并没有再提及“罢餐”二字。因此，我们将文本标题中是否出现了“罢餐”一词作为一个特征，如：“我们吃到活老鼠了!”北大学子大罢餐!

此外，通过观察 BBS 上的各种与“罢餐”相关的帖子，不难发现该类文本都具有文本长度较短的特点，因此也将该特点纳入至特征集，具体的实验设置为文本长度小于 500，特征设为 1，否则特征设为 0。

5 实验及结果分析

我们从网上收集到出现“罢餐”一词的文本共 454 篇，经过人工进行分类后，得到“罢餐”类文本 383 篇，“非罢餐”类文本 71 篇（分类标准如第 4 节所示）。将两类文本分别以 3：1 的比例分开，得到训练文本 342 篇，其中“罢餐”类 288 篇，“非罢餐”类 54 篇，其余 112 篇为测试文本。

运用最大熵分类器对训练语料进行训练，训练时所用的特征如 4.2 节所示。

实验所得的分类精度为 87.5%，即有 14 篇文本被分类错误，其中，包含 3 篇“罢餐”类文本，“罢餐”类文本召回率达到 96.8%，准确率为 89.3%。实验结果如下表所示

	自动分类为“罢餐”类	自动分类为“非罢餐”类
人工标记为“罢餐”类	92	11
人工标记为“非罢餐”类	3	6

从实验结果中可以看出，尽管分类精度并没有达到很高，但“罢餐”类文本的召回率高达 96.8%，已经满足了我们的预期。对“罢餐”类测试文本进行分析后，我们可以发现，在总共 95 篇文本中，有将近 70%的文本标题中包含“罢餐”一词。此外，我们可以很容易地从测试文本中找

出形如“抵制学校食堂菜价上涨”，“食堂的饭菜价格猛涨”，“大家这么团结”，“希望大家罢餐一日”等等这样包含 4.2 节中所列特征的句子。

综合以上分析，我们可以看出，运用情景语义学的框架进行特征提取的效果还是比较理想的，并且，基于这种方法提取的特征数目较少，因此运算比较简单，分类效率较高。

对于“非罢餐”类文本得到较低的召回率，究其原因在于，一是所有的文本都含有“罢餐”一词，因此，收集到的文本广义上都是罢餐相关的，这样的分类本身就存在难度。此外，“非罢餐”类文本只占所有文本的 15.6%，因此，在训练过程中，“非罢餐”类文本的特点并没有凸显出来。

6 结束语

基于情境框架的文本分类方法可以帮助我们有效地抓住文本所属类别的特征，不仅可以得到较高的召回率，而且大大降低了文档特征向量的维数，提高了分类效率。

在此次实验中，虽然“非罢餐”类文本的召回率较低，但“罢餐”类文本的召回率达到了 96.8%，鉴于我们重在找出在网络上具有“罢餐”倾向的文本，这样的实验结果初步达到了我们预期的效果。

在以后的工作中，为了提高文本分类的精度，不仅需要扩充语料的规模，也需要我们进一步研究基于认知情境的“罢餐”语义框架，通过不断挖掘该类文本的特征，进而得到更适于分类器分类的特征集合。

此外，在实验中，我们并没有对文本进行分词，原因在于对分词将大大降低分类器进行分类的速度，而且由于网络用语的不规范性，利用工具进行分词不能保证分词的正确性，进而导致特征的错误提取；此外，由于在特征词中出现了如“不去”这样的词，如将“不”与“去”作为两个特征词，会出现如下影响分类器进行分类的情况，即文本中分别出现了“不”与“去”这两个词，但这两个词并没有在文本中连续出现，因而不能体现“不去”这个特征要表达的真正意义。在以后的工作中，我们需要考虑在进行分类前，首先将文本进行分词，考察该种情况下是否会提高分类精度。

参考文献

[1]周茜，赵明生，扈曼，“中文文本分类中的特征选择研究”，中文信息学报，2004，18（3）：18-19

[2]高洁，吉根林，“文本分类技术研究”，计算机应用研究，2004，21（7）：18-21

[3]胡佳妮，徐蔚然，郭军，邓伟洪，“中文文本分类中的特征选择算法研究”，光通信研究，2005，（3）：44-46

[4]苏金树，张博锋，徐昕，“基于机器学习的文本分类技术研究进展”，软件学报，2006，17：1848-1859

[5]李荣陆，王建会，陈晓云，陶晓鹏，胡运发，“使用最大熵模型进行中文文本分类”，计算机研究与发展，2005，42（1）：94-101

[6]Winograd, Terry, *Language as a Cognitive Process*, Reading, Mass: Addison-Wesley Publishing Company, 1983

[7]袁毓林，“认知科学和汉语计算语言学”，《语言学前沿与汉语研究》，上海教育出版社，2005：171-198

[8]袁毓林，“基于认知的汉语计算语言学研究”。北京大学出版社，2008

[9]袁毓林、周强、陈振宇、张秀松、李湘、高嵩，“从认知假设到计算分析和程序实现---一种认知语言学研究的计算范式与技术路线”，将刊《当代语言学》，2009

[10]秦进，陈笑蓉，汪维家，陆汝占，“文本分类中的特征抽取”，2003，23（2）：45-46