DOI:10.13718/j.cnki.xsxb.2018.09.008

基于统计词典和特征加强的多语言文本分类①

龚　静[1]，　李英杰[1]，　黄欣阳[2]

1. 湖南环境生物职业技术学院 公共基础课部，湖南 衡阳，421005；
2. 南华大学 计算机学院，湖南 衡阳 421001

摘要：在统计双语词典的基础上，提出一种特征加强的多语言文本分类方法. 在执行文本分类时，考虑到其他语言的训练文本，使得多种语言的文本集合中均存在训练文本，放松了 MLTC 的要求. 特征加强是一种交叉检查过程，即获取两种语言所有特征的卡方统计后，通过语言中相关特征的辨识力，再次对语言的特征辨识力进行评估，以提高分类的可信度. 实验选择汉语或英语作为目标语言. 实验结果表明：提出的方法具有更高的分类精度，且对训练集规格的敏感度更低.

关　键　词：多语言文本分类；双语词典；特征加强；交叉检查；敏感度

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1000－5471(2018)09－0045－06

文本分类[1]一般是指根据文本内容分配文本到一个或多个预先定义类别中. 大多数分类方法是针对单一语言文本，但日常生活和工作会经常遇到多个语言文本，如跨国公司的某一部门会收到其他部门发来的多个语言文本；而很多官方语言也不止一种，如我国香港特区的官方语言是汉语和英语，加拿大的官方语言是法语和英语. 因此，多语言文本分类[2](multiple language text classification，MLTC)是一个非常值得研究的课题，其最大挑战是跨语言语义的互用性[3]. 多语言文本分类通常包含特征提取和选择、文本表示以及归纳[4]，各方法在特征提取和文本表示方面有所不同[5－9]. 虽然基于语料库的方法克服了机器翻译的限制，但它会受到目标任务中专属平行语料库的限制.

本文着重研究基于统计的双语词典. 提出了一种基于特征加强的多语言文本分类(feature enhanced-multiple language text classification，FE-MLTC)，其优势在于放松了现有 MLTC 的要求，即在多种语言中均存在训练文本，且特征加强使得分类效果更佳. 为了支持不同语言训练文本的可操作性，依靠基于统计的双语词典自动对一个平行文本集执行分类.

1　基于统计的双语词典生成

双语词典生成的步骤如图 1 所示. 该生成过程从提取和选择项开始，首先处理英语和汉语. 接着，对平行语料库[10]中子文档的每一个单词进行标记，从被标记的子文档中提取名词短语.

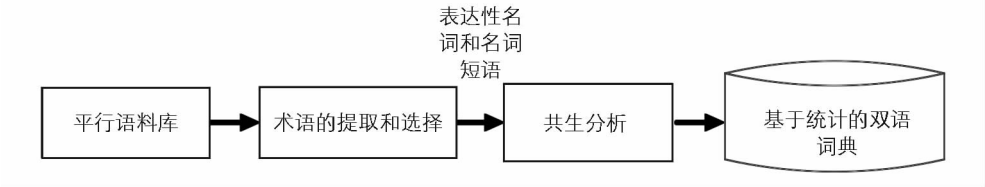

图 1　基于统计的双语词典生成过程

① 收稿日期：2016－11－22
基金项目：国家自然科学基金项目(60572137)；湖南省教育厅项目(12C1056；17C0599).
作者简介：龚　静(1972－)，女，教授，硕士，主要从事数据挖掘和分类算法等研究.

提取项后的选择项步骤会为每一个平行文件中的两种语言选出代表项. 本文采用 TF × IDF 机制，项 f_i(英语或中文) 在平行文件 d_i 中的项权值由如下公式计算获得：

$$tw_{ij} = tf_{ij} \times \log\left(\frac{N_p}{n_j}\right) \tag{1}$$

式中：tf_{ij} 是 f_j 在 d_i 中的项频率，N_p 是语料库中平行文件的总数，n_j 是 f_j 出现的平行语料库中子文档的数量. 在每个平行文档中，有着最高 TF × IDF 值的前 K_{dt} 个英语和汉语项同时出现在 δ_{DF} 文档中. 由于相关项通常同时出现在相同的平行文档中，通过共生分析测量 f_j 项和 f_h 项在平行文档 d_i 中的重要性权值(f_j 和 f_h 的语言相同)，具体如下：

$$cw_{ijh} = tf_{ijh} \times \log\left(\frac{N_p}{n_{jh}}\right) \tag{2}$$

式中：tf_{ijh} 是 d_i 中 tf_{ij} 和 tf_{ih} 中的最小量，n_{jh} 是 f_j 和 f_h 出现的平行文档的数量. f_j 和 f_h 间的相关权值：

$$rw_{jh} = \frac{\sum_{i=1}^{N_p} cw_{ijh}}{\sum_{i=1}^{N_p} tw_{ij}} \qquad rw_{kj} = \frac{\sum_{i=1}^{N_p} cw_{ijh}}{\sum_{i=1}^{N_p} tw_{ik}} \tag{3}$$

式中：rw_{jh} 阐释了 f_j 到 f_h 的相关权值，rw_{hj} 是 f_h 到 f_j 的相关权值. 如果某个语言中的项到另一个语言项的统计强度低于相关临界点 δ_{rw}，则移除此关联. 链接移除完成以后，从输入平行语料库中执行基于统计的双语词典，其中包含跨语言关联.

2 特征加强的多语言文本分类

MLTC 能够简单处理多种独立单一语言的文本分类. 然而，不能针对某种特定语言训练单一语言，本文提出一种基于双语词典支持的特征加强 FE-MLTC 法. 包括两个主要任务：①生成双语词典，即从平行语料库中构建基于统计的双语词典(针对英语和汉语)；②分类，即基于未分类的文档集，为每种语言生成一个文本分类模型. 提出的 FE-MLTC 算法核心如图 2 所示，当训练单一语言分类器时，本文还考虑了某种语言未分类的文档，以及基于统计的双语词典. 为了训练某种语言的单一语言分类器，分类任务分为以下 3 个步骤：特征提取、特征加强和选择以及文本表示.

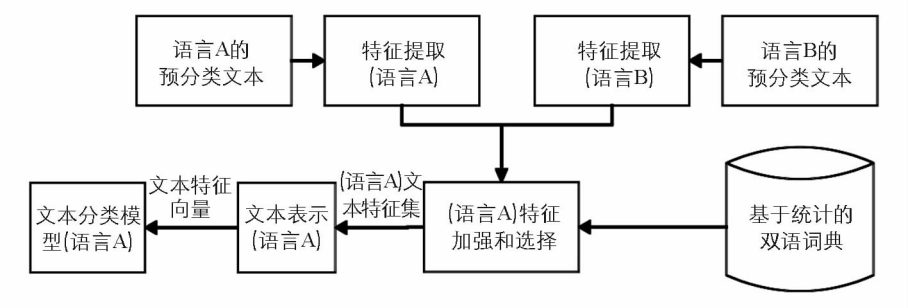

图 2 FE-MLTC 过程

2.1 特征提取

本文从两种语言的未分类文本中提取特征. 使用基于规则的词性标签和名词短语分析程序来提取名词和名词短语作为未分类英语文本的特征. 对于未分类的中文文本，采用基于词典和基于统计的混合方法从训练语料库中提取中文项[11].

2.2 特征加强和选择

该小节是本文的核心部分. 特征提取过程中，本文首先评估了每个特征在其各自的训练语料库和语言中的辨识力，评估采用了卡方统计. 当测量特征 f_j 和类别 C_i 独立时，卡方统计值趋近于 0. 这里采用 f_j 和 C_i 的双向预防表，设定 n_{r+} 是拥有 f_j 的 C_i 类别的文本数量，n_{r-} 是 C_i 中 f_j 未出现的文本数量，n_{n+} 是 f_j 出

现的除 C_i 类别的文本数量，n_{n-} 是 f_j 未出现的除 C_i 类别的文本数量，n 是训练语料库的文本总数. 因此，与 C_i 相关的 f_j 的 χ^2 统计如下：

$$\chi^2(f_j, C_i) = \frac{n \times (n_{r-} + n_{n-} - n_{r-} n_{r+})^2}{(n_{r+} + n_{r-})(n_{n+} + n_{n-})(n_{r+} + n_{n+})(n_{r-} + n_{n-})} \tag{4}$$

卡方统计之后，根据权重平均机制计算所有类别 T 中特征 f_j 的完整 χ^2 统计：

$$\chi^2(f_j) = \sum_{C_i \in T} p(C_i) \cdot \chi^2(f_j, C_i) \tag{5}$$

式中 $p(C_i)$ 是分类的 C_i 文档数量.

在通过卡方统计获取了两种语言的所有特征后，考虑其他语言中相关特征的辨识力，再一次对一种语言的特征辨识力进行评估. 这种在两种语言中交叉检查的原因在于：如果一种语言的一个特征及其相关特征的卡方分数较高，则很可能该特征有更高的辨识力. 然而，两种不同的评估结果(即某个特征在某种语言中的卡方值很高，但是与该特征紧密联系的特征在其他语言中的卡方值很低)降低了特征辨识力的可信度. 本文将这种交叉检查过程作为特征加强.

假设语言 L_1 中所有的 N_1 特征都提取于 L_1 中未分类的文本，L_2 中所有的 N_2 特征都提取于 L_2 中未分类文本. 鉴于 L_1 中的特征 f_j，根据之前生成的基于统计的双语词典，将 $R(f_j)$ 设为 L_2 中的特征集，且其对 f_j 有着直接跨语言的关联. 从其在 L_2 中的相关特征中提取 L_1 中 f_j 的联合权值，具体如下：

$$aw(f_j) = \begin{cases} \dfrac{\sum\limits_{\forall g_h \in R(f_j)} \chi^2(g_h) \times rw_{g_h f_j}}{|R(f_j)|} & \text{如果 } |R(f_j)| \neq 0 \\ 0 & \text{如果 } |R(f_j)| = 0 \end{cases} \tag{6}$$

式中：$\chi^2(g_h)$ 是特征 g_h 的卡方值，$rw_{g_h f_j}$ 是从 g_h 到 f_j 的相关权值. (6)式既考虑了相关特征在其它语言(如 $\chi^2(g_h)$)中的辨识力，还考虑到了它们与 f_j 的相关权值. 公式(6)考虑了所有关联的特征，并用相关特征的数量将其规范化. 权值公式如下：

$$aw(f_j) = \begin{cases} \dfrac{\sum\limits_{\forall g_h \in R(f_j)} \chi^2(g_h) \times rw_{g_h f_j}}{|R(f_j)|} \cdot \log \dfrac{N_2}{|R(f_j)|} & \text{如果 } |R(f_j)| \neq 0 \\ 0 & \text{如果 } |R(f_j)| = 0 \end{cases} \tag{7}$$

式中：N_2 表示从 L_2 未分类的训练文本中提取的特征数量；$\log\left(\dfrac{N_2}{R(f_j)}\right)$是逆项频率(ITF). 其中，过度频繁的项 ITF 比特殊项(即与其他语言关联少的项)的值要小. 因此，该权值公式更适用于特殊项，而不是过度频繁的项. 因此，本文结合两种语言中训练文本的特征 f_j 的权值，使用如下公式得到了特征 f_j 的全部权值

$$w(f_j) = \chi^2(f_j) \times aw(f_j)^{\alpha} \tag{8}$$

其中 $\alpha(\alpha \geqslant 0)$ 表示特征 f_j 在目标语言的卡方统计分数和其他语言特征 f_j 的权值平衡.

当 $\alpha = 0$ 时，特征整体权值的评估完全依赖于目标语言. 相比之下，当 α 增加时，特征的评估更多地依赖于其他语言中的相关特征. 因为 $aw(f_j)$ 是其他语言中相关特征的卡方统计值的平均.

在获得两种语言中所有特征的全部权值后，执行了特征选择. 针对每一种语言(L_1 或 L_2)，本文选择有着最高整体权值的特征来表示各自语言的每一个训练文本.

2.3　文本表示

该步骤中，每一种语言的训练文本由先前选择出的相应特征集表示. 由于文本相对短小，故选择二进制来表示文本. 因此，每一个训练文本对应的特征向量 $\boldsymbol{d}_i$ 如下：

$$\boldsymbol{d}_i = (w_{i1}, w_{i2}, \cdots, w_{ik}) \tag{9}$$

式中：k 是先前选择出的特征数量；w_{ij} 表示 $\boldsymbol{d}_i$ 中特征 f_j 的状态，用 1 表示存在，用 0 表示消失.

3　实验评估

3.1　数据集

如上所述，基于统计的双语词典的构建需要两种语言的平行文本. 实验收集我国香港特别行政区的政

府信息中心发布的新闻(http://www.info.gov.hk/),以建造基于统计的双语词典.收集的平行语料库包含7 689则中英文对照新闻.

本文还收集了两种额外的单一语言文本语料库以评估提出的FE-MLTC效果.这些中英语料库包括从香港政府信息中心收集的新闻.手动将收集的新闻文本分到8个类别中(贸易与经济、通信与IT、文化与休闲、教育、健康与环境、住宅与土地、安全、交通运输).为了避免受不同文本类别大小的影响,本文在每个语料库的每个类别中随机选择了相同数量的新闻文本.中英语料库中每个类别都包含了76个新闻文本,每个语料库中新闻文本的总数达650个.为了评估,将这两个语料库融合到一个多语言语料库中.

本文随机在中英语料库中选择50%的文本作为训练数据集,并将每个语料库中剩下的文本作为测试数据集.为避免随机抽样带来的影响,随机抽样并训练-测试30次,将30次评估结果的平均值作为最终结果.

3.2 实验结果比较

在双语词典的构建过程中,选择出的项必须满足每一种语言中平行文本的频率临界值δ_{DF},确定文本的前K_{dt}项.为避免可能出现的过拟合问题,参照文献[9],设置δ_{DF}取3,K_{dt}取30,δ_{rw}取0.15.

特征数量k和特征加强步骤α对实验的影响很大,因此,本文特别讨论了k值和α值,其中,k取值200到2000,α取值范围为[1,3].比较的方法有文献[7]的跨语言文本分类方法SAAW和文献[9]分类方法CLTC.实验使用的分类器有采用高斯核函数的朴素贝叶斯分类器和支持向量机(supported vector machine,SVM),目标语言的不同也会对实验产生一定影响,故这里考虑目标语言为英文和目标语言为中文两种情况.

3.2.1 目标语言为英语

特征数k与分类精度的关系图如图3(a)所示.当k值从250上升时,CLTC[9]的分类精度最高,其次是SAAW[7].3种α值情况下,FE-MLTC都呈现出与k值相似的模式.当α取1时,如果k值由200变化到800,FE-MLTC分类的准确性会提升,当k值达到或超过1000时,准确性会下降.整体来看,在使用朴素贝叶斯分类时,当α值为2,k值为1200时,FE-MLTC法的效果最好.

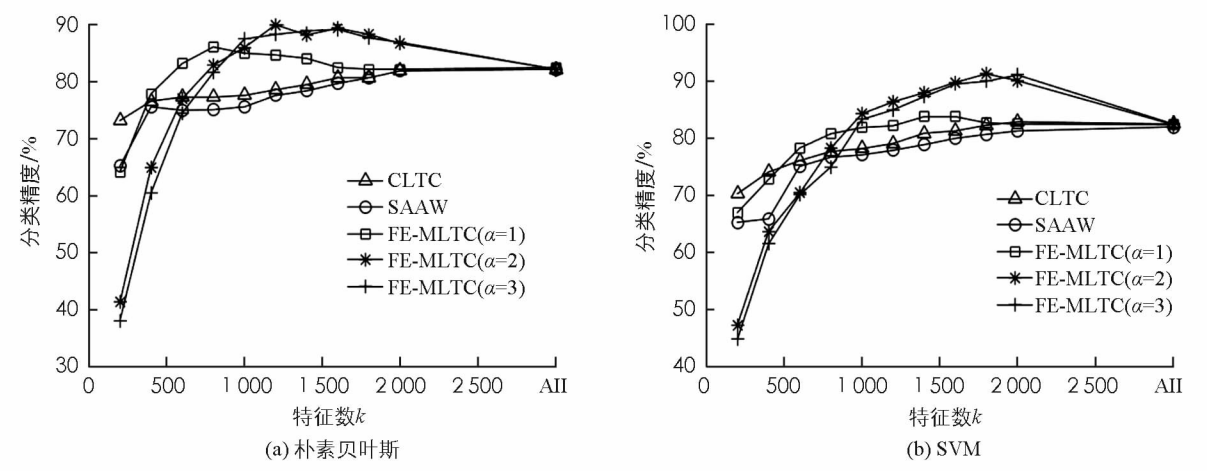

图3 特征数k与分类精度的关系图

当分类器采用SVM时,FE-MLTC法中的k和α值的效果与采用朴素贝叶斯算法的结果相似.具体如图3(b)所示,当k逐渐上升的时候,CLTC[9]和SAAW[7]的分类精度都有所增加,另一方面,当α值为2,k值为1800时,FE_PLTC法的效果最好.

虽然当α等于2,k大于或等于800时,提出的FE_MLTC法在两种有监督的学习方法中整体表现优于CLTC和SAAW,但是当k低于600时,它的效果相对较差.一个可能的原因是,FE_MLTC选择出的特征可能是文本分类的代表,但是与CLTC法选择出来的特征相比,并不经常出现在测试数据集中.例如,当$\alpha=2$,$k=200$时,9.72%的测试文本并没有包含FE_MLTC法所选择出来的特征.然而,CLTC法中只有0.69的测试文本遇到了同样的问题.当k值升至600时,FE_MLTC法选择的特征不包含在测试文档中的百分比迅速降低(从9.72%到0.35%).因此,FE_MLTC需要大量特征.

3.2.2　目标语言为汉语

目标语言为汉语时，分类也使用有监督的方法，其结果如图 4 所示. 当 k 从 200 升到 1 800 时，使用朴素贝叶斯的 CLTC 和 SAAW 的分类准确性均有所提高；当 k 超过 1 800 时，分类精度开始下降. 相反，当 $\alpha=3$，$k=1000$ 时，FE_MLTC 法的分类准确性最高. 使用 SVM 分类，在特征选择最少时，CLTC 效果最好，随着特征数量增加，$\alpha=3$，$k=1200$ 时，FE_MLTC 法的分类效果最好.

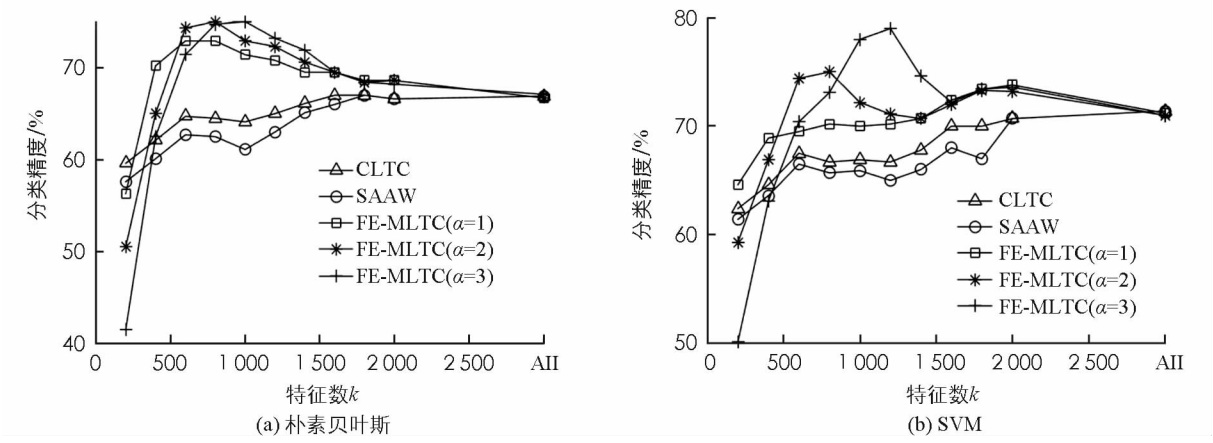

图 4　特征数 k 与分类精度的关系

综上，在目标语言为英语和汉语的两种情况下，使用英文作为目标语言的收益稍微高点，这可能是由于英文的特征向量表示方式比汉语更优. 对于本文，朴素贝叶斯比 SVM 更好.

3.2.3　汉明损失和首位误差评价

下面使用汉明损失和首位误差两种目前较为流行的标准进行评价[3]. 在汉明损失和首位误差中，分值越小，分类表现越好.

表 1－2 给出了各数据集上的比较结果，数据集为 8 个类别的新闻文本. 在汉明损失中，本文方法在大多数情况下都获得了更好的分类效果，总体来说本文方法更优. 表 2 的首位误差的结果显示：依赖翻译质量的 SAAW 没有平行语库 CLTC 和本文方法的分类效果好. 将分类问题转化为依靠翻译质量的标签问题，造成了首位排序的误差较大.

表 1　分类方法的汉明损失比较

数据集	SAAW[7]	CLTC[9]	FE-MLTC
贸易与经济	0.061 3	0.050 4	0.052 9
通信与 IT	0.040 1	0.034 9	0.033 7
文化与休闲	0.049 9	0.038 9	0.037 6
教育	0.059 8	0.058 8	0.055 2
健康与环境	0.032 1	0.022 1	0.026 7
住宅与土地	0.031 2	0.025 6	0.021 9
安全	0.039 6	0.034 2	0.033 9
交通运输	0.035 8	0.039 1	0.033 4

表 2　各分类方法的首位误差比较

数据集	SAAW[7]	CLTC[9]	FE-MLTC
贸易与经济	0.598	0.477	0.422
通信与 IT	0.604	0.236	0.127
文化与休闲	0.630	0.340	0.456
教育	0.601	0.201	0.237
健康与环境	0.507	0.388	0.301
住宅与土地	0.499	0.284	0.280
安全	0.402	0.188	0.156
交通运输	0.224	0.241	0.187

4　结论与展望

本文提出了一种基于特征加强的多语言文本分类方法，该方法放松了现有 MLTC 的要求，使得多种语言中均可以存在训练文件. 提出的特征加强机制更好地执行了文本分类的任务，将未分类的多语言文本集作为训练文本来构建文本分类模型，提出的方法有望运用到事件追踪、垃圾邮件过滤、B2C 电子商务等.

由于当前评估所用的文本长度比较短，而文本的长短会影响到文本表示机制的选择等问题，未来会考虑更长的文本. 另外，两种以上的语言文本也需要进一步研究.

参考文献:

[1] 赖 娟,金 澎,洪艳伟. 文本分类中的主动多域学习[J]. 西南师范大学学报(自然科学版),2014,39(7):108-114.

[2] 罗远胜,王明文,勒中坚,等. 双语潜在语义对应分析及在跨语言文本分类中的应用研究[J]. 情报学报,2013,32(1):86-96.

[3] 刘志红. 多语种多类别体系下文本自动分类系统的研究与实现[D]. 沈阳:东北大学,2010.

[4] FORTUNA B, DEMEESTER T, DEVELDER C. Towards Large-scale Event Detection and Extraction from News[C]// The Workshop on Large-Scale Online Learning & Decision Making. New York: IEEE Press, 2014: 1-3.

[5] PRETTENHOFER P, STEIN B. Cross-language Text Classification Using Structural Correspondence Learning[C]// ACL 2010, Meeting of the Association for Computational Linguistics. New York: IEEE Press, 2010: 1118-1127.

[6] 张金鹏,周兰江,线岩团,等. 基于跨语言语料的汉泰词分布表示[J]. 计算机工程与科学,2015,37(12):2358-2365.

[7] 张玲玲,冀俊忠,贝 飞,等. 基于句法分析和属性概率权重的跨语言情感分类算法[J]. 模式识别与人工智能,2015,28(11):1002-1012.

[8] NI X, SUN J T, HU J, et al. Cross Lingual Text classification by Mining Multilingual Topics from Wikipedia[C]// Forth International Conference on Web Search and Web Data Mining, WSDM 2011. New York: IEEE press, 2011: 375-384.

[9] WEI C P, LIN Y T, YANG C C. Cross-lingual Text Categorization: Conquering Language Boundaries in Globalized Environments[J]. Information Processing & Management, 2011, 47(5): 786-804.

[10] 熊文新. Web、语料库与双语平行语料库的建设[J]. 图书情报工作,2013,57(10):128-135.

[11] 司 莉,庄晓喆,贾 欢. 近10年来国外多语言信息组织与检索研究进展与启示[J]. 中国图书馆学报,2015,34(4):112-126.

Multiple Language Text Classification Method Based on Statistical Dictionary and Feature Enhancing

GONG Jing[1], LI Ying-jie[1], HUANG Xin-yang[2]

1. *Department of Public Basic Course, Hunan Polytechnic of Environment and Biology, Hengyang Hunan 421005, China;*

2. *Computer School, University of South China, Hengyang Hunan 421001, China*

Abstract: Aiming at the problem that multiple language text classification (MLTC) can only solve single language text classification problem of multiple independent, on the basic of statistical bilingual dictionary, multiple language text classification based on feature enhancing has been proposed. In the implementation of text classification, the training texts of other languages have been taken into account, which makes the text of a variety of languages in the training texts. And it relaxes MLTC requirements. Feature enhancing is a processing of cross examination. After chi square statistics of all the features for the two languages is obtained, the identification of language feature is reassessed through the feature identification to improve the reliability of classification. Chinese or English is chosen as the target language in the experiment. Experimental results show that the proposed method has a higher classification accuracy, and the sensitivity of the training set is lower.

Key words: multiple language text classification; bilingual dictionary; feature enhancing; cross examination; sensitivity

责任编辑 张 枸