第25卷 第4期
2011年7月
中文信息学报
JOURNAL OF CHINESE INFORMATION PROCESSING
Vol. 25, No. 4
Jul., 2011

**文章编号：** 1003-0077(2011)04-0039-06

# 最大熵和规则相结合的藏文句子边界识别方法

李 响[1]，才藏太[2]，姜文斌[1]，吕雅娟[1]，刘 群[1]

（1. 中国科学院 计算技术研究所，中国科学院 智能信息处理重点实验室，北京 100190；
2. 青海师范大学 计算机学院，青海 西宁 810008）

**摘 要：** 句子边界识别是藏文信息处理领域中一项重要的基础性工作，该文提出了一种基于最大熵和规则相结合的方法识别藏语句子边界。首先，利用藏语边界词表识别歧义的句子边界，最后采用最大熵模型识别规则无法识别的歧义句子边界。该方法有效利用藏语句子边界规则减少了最大熵模型因训练语料稀疏或低劣而导致对句子边界的误判。实验表明，该文提出的方法具有较好的性能，F1 值可达 97.78%。

**关键词：** 最大熵；句子边界识别；藏文信息处理

**中图分类号：** TP391　　　**文献标识码：** A

# A Maximum Entropy and Rules Approach to Identifying Tibetan Sentence Boundaries

LI Xiang[1], CAI Zangtai[2], JIANG Wenbin[1], LV Yajuan[1], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China;
2. Computer College, Qinghai Normal University, Xining, Qinghai 810008, China)

**Abstract:** Sentence boundary identification is a fundamental work in the field of Tibetan information processing. This paper proposes a maximum entropy and rules approach to identifying Tibetan sentence boundaries. First, the Tibetan boundary vocabulary based detector identifies the ambiguous sentence boundaries. Second, the maximum entropy model based detector identifies the ambiguous sentence boundaries which the former detector can't identify. By making use of Tibetan sentence boundary rules, this approach further reduces the number of the incorrect sentence boundary identified by maximum entropy model owing to the sparse and inferior training corpus. The experiments show that this approach has a good performance in terms of 97.78% F1-measure.

**Key words:** maximum entropy; sentence boundary identification; Tibetan information processing

## 1 引言

藏语是属于汉藏语系的一种古老语言，在漫长的语言演变过程中，藏语形成了独特的标点符号体系，并仍然在现代藏语文本中得到广泛使用。

藏文标点符号体系仅含有限的标点符号，并且标识句子结束的标点符号存在较多的歧义，功能不确定，严重影响了藏语句子边界的准确识别。作为藏文信息处理的一项基础性工作和藏语自然语言处理的一项关键技术，藏语句子边界识别问题解决的好坏直接影响到词性标注，词语切分、句法分析及机器翻译等其他藏文自然语言处理应用的性能，因此，解决现代藏语句子边界的自动识别问题显得日益重要。

现有的藏语句子边界识别方法主要以采用规则方法为主[1]，可以对特定领域的藏文文本实现较好的识别准确率，但是该方法需要制作针对性的规则，人工代价较大，同时领域适应性较差。本文提出了一种最大熵和规则相结合的藏语句子边界识别方

---

**收稿日期：** 2011-04-20　**定稿日期：** 2011-05-21

**基金项目：** 国家自然科学基金重大研究计划培育项目(90920004)；国家自然科学基金重点资助项目(60736014)

**作者简介：** 李响(1987—)，男，硕士生，研究方向为统计机器翻译；才藏太(1974—)，男，副教授，研究方向为藏文信息处理；姜文斌(1984—)，男，博士，研究方向为统计机器翻译。

法。最大熵模型可以增强对不同领域文本句子边界识别的鲁棒性，但由于最大熵模型会因训练语料稀疏或低劣而产生对句子边界的错误识别，而引入规则方法可以提高最大熵模型的准确率。本文提出的方法首先利用人工总结和分析藏语语料得到的藏语句子边界词表和非边界词表识别藏语句子边界，对于无法识别的藏语句子边界，最后利用最大熵模型进行识别。实验表明，本文提出的方法可以对现代书面藏语文本的句子边界进行较好的识别。

本文包括如下部分：第2节首先介绍藏语标点符号体系和对藏语句尾的分析，提出基于规则方法的藏语句子边界识别方案；第3节主要说明最大熵模型的原理，提出基于最大熵方法的藏语句子边界识别方案；第4节为实验结果；第5节为总结和展望。

## 2 规则方法

### 2.1 藏语标点符号体系

藏语独特的标点符号系统中涉及标识句子结束的标点符号主要有[2]：

- 音节点“་”：藏文是拼音文字，以音节为单位，各音节间用因节点划分音节关系；
- 楔形符“།”：用来表示藏文的词、句子的停顿或结束；
- 云头号“༄༅།”：多用于文章篇首，表示文字的开始。

在这些标点符号中，楔形符在藏文句子中主要用于句末，也可用于词或者短语之后，在功能上相当于汉语标点符号的顿号、逗号、句号、问号及叹号。例如：

དོན་ཚན་གཉིས་པ།སྤྲུལ་སྐུ་ཀྱི་ཡང་སྲིད་ངོས་འཛིན་བྱེད་སྐབས་རྒྱལ་ཁབ་གཅིག་གྱུར་སྲུང་སྐྱོང་དང་།མི་རིགས་མཐུན་སྒྲིལ་སྲུང་སྐྱོང་།ཆོས་ལུགས་འཆམ་མཐུན་དང་སྤྱི་ཚོགས་འཆམ་མཐུན་སྲུང་སྐྱོང་།བོད་བརྒྱུད་ནང་བསྟན་གྱི་རྒྱུན་ལྡན་སྒྲིག་སྲོལ་སྲུང་སྐྱོང་བཅས་བྱ་རྒྱུའི་རྩ་དོན་ལ་བརྩི་སྲུང་ཞུ་དགོས།སྤྲུལ་སྐུ་ཀྱི་ཡང་སྲིད་ངོས་འཛིན་བྱེད་སྐབས་བོད་བརྒྱུད་ནང་བསྟན་གྱི་ཆོས་ལུགས་སྲོལ་དང་ལོ་རྒྱུས་ཀྱི་ངེས་སྲོལ་ལ་བརྩི་འཇོག་ཞུ་དགོས།འོན་ཀྱང་ཆེས་མེད་དུ་བཏང་ཟིན་པའི་བཀས་བསྐོད་རྒྱུད་འཛིན་གྱི་དམིགས་བསལ་དབང་ཆ་སླར་གསོ་བྱ་མི་ཆོག

第二条 活佛转世应当遵循维护国家统一、维护民族团结、维护宗教和睦与社会和谐、维护藏传佛教正常秩序的原则。活佛转世尊重藏传佛教宗教仪轨和历史定制，但不得恢复已被废除的封建特权。

从上面的例句中我们可以通过逗号、顿号和句号的位置准确识别出汉语句子边界，但是对应的藏语句子共出现了6个楔形符，导致无法准确识别藏语句子边界。

### 2.2 藏语句尾结构分析

藏语的句子语序结构属于SOV型，即{主语＋宾语＋谓语}的语序结构。藏语句子的构成是以动词为核心，运用各种关系词将词语联接起来组成句子的过程。动词的句法位置始终位于句子结尾部分，因此藏语句子中谓语部分的末端应当是整个句子的结尾。但是单独分析谓语部分时，其谓语结构又各有不同的特点，构成要素之间相互影响。一般在藏语句子的谓语部分中核心动词后总是附加包含有一些其他成分，例如辅助动词等，其谓语的语序格式为：{（谓语动词（＋状语补语）（＋助动词[情态和趋向]）（＋体貌－示证标记）（语气词）}[3]。

另外，藏语句子的主要成分一般都要与格助词相关联。格关系是动词和其周围对象发生事件的约束关系，只有这样才能把句子各成分之间的语义关系表达清楚[4]。通过对48 000句藏文语料中歧义句子边界左侧的格助词等词的分析和统计，我们从中总结出406个非边界词和166个边界词，这些词可以确定性地表示藏语句子边界的功能，如表1和表2所示。

**表1 非边界词表(部分)**

| | | | | |
|---|---|---|---|---|
| ༄༅། | ་ཅིང་། | ་གདམ། | ་གནད། | ་གཙང་། |
| ་གར། | ་གཞན། | ་གཅོད། | ་གཉིས། | ་གཉེར། |

**表2 边界词表(部分)**

| | | | | |
|---|---|---|---|---|
| ་མིན། | ་སྲིད། | ་ཆོག | ་བྱུང་། | ་གཏོང་། |
| ་རུང་། | ་ཐུབ། | ་བཏང་། | ་བྱེད། | ་རྒྱུ། |

基于规则的方法利用表1和表2的词表识别歧义的藏语句子边界，对于每一次的识别结果，标记该句子边界的确定功能，从而根据该标记决定最大熵模型需要识别的未确定的句子边界。

## 3 最大熵方法

现有许多机器学习方法可以应用于句子边界识别问题，例如，决策树[5]、神经网络[6]和条件随机

场[7]等。由于最大熵模型已经非常成熟，可以采用那些可以开源的最大熵训练工具包来进行训练，因此本文选择最大熵模型来解决藏语句子边界识别问题。

### 3.1　最大熵原理

如果将一段文本看作一个词序列，则可将句子边界识别问题视为一个将文本划分为句子的随机过程，建立随机过程的联合概率分布模型 $p$，$p\in P$，输出值集合 $Y=\{\mathrm{sb},\mathrm{nsb}\}$，$y\in Y$，其中 $y$ 是歧义句子边界是否为有效边界的结果，在这个随机过程中，$Y$ 受到上下文信息 $x$ 的影响，上下文集合 $X$，$x\in X$，其中 $x$ 表示此序列中所有可能的上下文特征组合，同时，从训练数据中获得 $N$ 个样本的集合 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 $(x_1,y_1)$ 是观察到的一个事件，我们可以根据这些样本定义一个事件空间 $X\times Y$，而对于句子边界识别问题，特征是一个二值函数 $f:X\times Y\to\{0,1\}$。

对于一个特征 $(x_0,y_0)$，定义特征函数如下：

$$f(x,y)=\begin{cases}1 & \text{如果 } y=y_0\text{，并且 } x=x_0\\ 0 & \text{其他情况}\end{cases}\tag{1}$$

对于一个特征 $(x_0,y_0)$，在样本中的期望值如下：

$$\bar{p}(f)=\sum_{(x_i,y_i)}\bar{p}(x,y)f(x,y)\tag{2}$$

其中 $\bar{p}(x,y)$ 是 $(x,y)$ 在样本中出现的概率。

对于一个特征 $(x_0,y_0)$，在模型中的期望值如下：

$$p(f)=\sum_{(x_i,y_i)}p(x_i,y_i)f(x_i,y_i)\tag{3}$$

最大熵模型的约束条件为对每一个特征 $(x,y)$，模型所建立的条件概率分布的特征期望值应与从训练样本中得到特征的样本期望值一致，如公式：

$$p(f)=\bar{p}(f)\tag{4}$$

联合概率分布模型 $p$ 的熵函数如公式：

$$H(p)=-\sum p(x,y)\log p(x,y)\tag{5}$$

最大熵模型如公式：

$$p^{*}=\arg\max_{p\in C}H(p)\tag{6}$$

其中，$C$ 是满足条件约束的模型集合，下面需要寻找 $p^{*}$，$p^{*}$ 具有如下的形式：

$$p^{*}(y\mid x)=\frac{1}{Z(x)}\exp\Big(\sum_i\lambda_i f_i(x,y)\Big)\tag{7}$$

其中，$Z(x)$ 是归一化常数，表示形式如下：

$$Z(x)=\sum_y\exp\Big(\sum_i\lambda_i f_i(x,y)\Big)\tag{8}$$

$\lambda_i$ 是模型参数，每一个特征 $f_i$ 对应一个 $\lambda_i$，$\lambda_i$ 决定了每个特征 $f_i$ 对概率分布的贡献程度，同时，可以采用 GIS 算法[8]对这些模型参数进行参数估计。

### 3.2　特征选择

针对藏语句子边界识别问题，选择有效的句子边界特征是使用最大熵模型需要解决的一个关键问题。根据藏语句子边界上下文的特点确定模型的上下文激发环境，从而选择所需特征。本文考察了影响藏语句子边界识别的多种因素，定义了藏语句子边界识别的特征模板，具体的特征模板如表 3 所示，其中单词表示藏语中两个音节点之间的字串。

表 3　藏语句子边界识别的特征模板

| 原子特征模板 | 模板意义 |
|---|---|
| L2L1 | 边界左边第 2 个单词和第 1 个单词 |
| L2 | 边界左边第 2 个单词 |
| L1 | 边界左边第 1 个单词 |
| L1Len | 边界左边第 1 个单词长度 |
| L1Syl | 边界左边第 1 个单词的末尾字 |
| R1 | 边界右边第 1 个单词 |
| R1Len | 边界右边第 1 个单词长度 |
| R2 | 边界右边第 2 个单词 |
| R1R2 | 边界右边第 1 个单词和第 2 个单词 |

同时，由于训练语料存在较多的数字、标点和英文等非藏文字符，为了避免数据稀疏对藏语句子边界识别效果的影响，我们采用泛化的方法，对这些字符进行分类处理，形成了如表 4 所示的单词类型表。

表 4　单词类型表

| 单词类型 | 单词类型意义 |
|---|---|
| digit | 阿拉伯数字 |
| lpunct | 左标点符号，包括(，[，<，{，《，"，' |
| rpunct | 右标点符号，包括}，]，>，》，"，' |
| cpunct | 除左、右符号外的其他标点符号 |
| tword | 藏文 |
| abc | 英文 |

下面通过图 1 来简要说明对 2.1 节中例句边界特征模板的抽取，其中采用简单的特征集合{L2，L1，L1Len，R1}。如图 1 所示，将每一个句子边界

的功能标记为断句(sb)或不断句(nsb)，用左斜线划分藏语单词，空格表示非有效句子边界，@符号表示有效句子边界，共抽取了5个边界特征。

@/བློ/སྤྲུལ/གྱི/ཡང/སྲིད/ངོས/འཛིན/བྱེད/སྐབས/རྒྱལ/ཁབ/གཅིག/གྱུར/སྲུང/སྐྱོང/དང/ /མི/རིགས/མཐུན/སྒྲིལ/སྲུང
/སྐྱོང/ /ཆོས/ལུགས/འཆམ/མཐུན/དང/སྤྱི/ཚོགས/འཆམ/མཐུན/སྲུང/སྐྱོང/ /བོད/བརྒྱུད/ནང/བསྟན/གྱི/རྒྱུན/ལྡན
/སྒྲིག/སྲོལ/སྲུང/སྐྱོང/བཅས/བྱ/རྒྱུའི/རྩ/དོན/ལ/བརྩི/སྲུང/ཞུ/དགོས/@

| y | x=L2 | x=L1 | x=L1Len | x=R1 |
|---|---|---|---|---|
| sb | སྐྱོང | དང | 2 | མི |
| sb | སྲུང | སྐྱོང | 2 | ཆོས |
| sb | སྲུང | སྐྱོང | 2 | བོད |
| nsb | ཞུ | དགོས | 3 | ཨ |

图1 藏语句子边界特征模板集合的抽取

## 4 实验

实验采用的藏语训练语料规模为48 000句，测试语料规模为140句，对应的参考语料规模为560句，测试语料和参考语料的句数比为1∶4，从而可以较客观地测试句子边界识别性能。

本文采用了张乐开发的最大熵模型训练工具包①，在训练过程中，迭代次数设为100次，为了避免过训练，高斯先验设为1.0，其他的参数都为缺省设置。

### 4.1 评价指标

为了客观评价本文提出的藏语句子边界识别方法的性能，依据本文提出的方法，我们实现了一个藏语句子边界自动识别系统，以准确率、召回率和F1值为指标对系统的藏语句子边界识别结果进行评价，相关计算公式如下所示。

$$\text{准确率}\ R = \frac{\text{系统正确识别的藏语句数}}{\text{参考语料的藏语句数}} \quad (9)$$

$$\text{准确率}\ P = \frac{\text{系统正确识别的藏语句数}}{\text{系统识别的藏语句数}} \quad (10)$$

$$F1 = \frac{2 \times R \times P}{R + P} \quad (11)$$

对于基于最大熵模型解决藏语句子边界识别问题，当需要对藏文文本识别句子边界时，利用公式(7)可以获得句子边界标记的概率，而句子边界识别可以看作两类情况的分类问题，因此，实验采用$p(y=sb|x) \geqslant 0.5$作为判别句子边界的阈值。

### 4.2 特征模板的选择

由于特征模板可以形成很多特征集合，但在对藏语句子边界进行充分分析的情况下，没有必要尝试所有的特征集合，根据经验和分析，在表3描述的特征模板的基础上选择特征形成如下6个特征模板集合。

(1) 特征集合A：A＝{L2L1, L2, L1Syl, L1Len, L1, R1, R1Len, R1Syl, R2, R1R2}，包含了表3中的所有特征模板，作为评估其他特征模板的参考。

(2) 特征集合B：B＝{L1Len, L1}，用于评价句子边界左侧第一个单词及单词长度对句子边界识别的影响。

(3) 特征集合C：C＝{L1Syl, L1Len, L1}，用于评价句子边界左侧第一个单词的尾部字对句子边界识别的影响。

(4) 特征集合D：D＝{L2, L1Syl, L1Len, L1}，用于评价句子边界左侧第二个单词对句子边界识别的影响。

(5) 特征集合E：E＝{L2L1, L2, L1Syl, L1Len, L1}，用于评价句子边界左侧第二个单词和第一个单词共现对句子边界识别的影响及句子边界左侧特征对句子边界识别的贡献。

(6) 特征集合F：F＝{L2L1, L2, L1}，用于评价候选句子边界左侧第二个单词和第一个单词共现，第二个单词以及第一个单词这种简单特征模板

---

① 网址：http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

集合对句子边界识别的影响。

为了验证每个特征集合的性能以及选择最有效的特征集合，分别采用以上每个特征集合识别藏语句子边界，实验结果如图 2 所示。

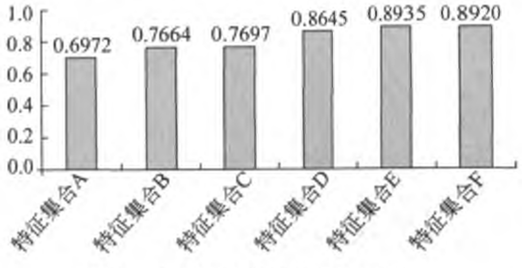

图 2　6 个特征集合的测试结果

实验结果表明，采用特征集合 *E* 的实验结果最好，特征集合 *A* 的实验结果最差，因此选择特征集合 *E* 作为最有效的特征模板集合。另外，从实验结果比较可见，特征集合 *A* 的实验结果最差，说明句子边界右侧特征并不能对实验结果产生较大贡献，所含信息量较少。

### 4.3　不同识别方法的比较

为了验证不同方法对句子边界识别性能的影响，分别采用基于规则的方法、基于最大熵模型的方法以及最大熵模型和规则相结合的方法，对相同测试语料测试藏语句子边界识别性能，其中特征集合采用 4.2 节中的特征集合 *E*，实验结果如表 5 所示。

表 5　不同边界识别方法的实验结果

| 方　法 | 准确率 | 召回率 | F1 值 |
|---|---|---|---|
| 规则方法 | 0.300 7 | 0.076 8 | 0.122 3 |
| 最大熵模型方法 | 0.873 7 | 0.914 3 | 0.893 5 |
| 规则＋最大熵模型方法 | 0.975 1 | 0.980 4 | 0.977 8 |

表 5 表明，虽然最大熵模型方法已经实现了较好的性能，但是小规模训练语料并不能较好地反映藏语复杂的边界特征分布，通过结合藏语句子边界的规则，藏语句子边界识别性能得到大幅提高，减少了最大熵模型对歧义句子边界识别的误判，具有对错误识别较好的约束作用。

## 5　总结与展望

藏文标点符号的特殊性和复杂性使我们不易准确地识别藏语句子边界，从而影响其他藏文自然语言处理的相关工作。通过对藏文语料以及对藏语句尾结构的分析，结合藏文语法规则，本文总结出大量的边界词及非边界词，可以利用词表在一定程度上确定歧义的藏语句子边界的功能，而对于规则不能识别的藏语句子边界，采用最大熵模型进行边界识别。实验结果表明，本文提出的最大熵与规则相结合的藏语句子边界识别方法能够较好的解决藏语句子边界识别问题。

在此基础上，我们一方面计划扩大训练语料，减少数据稀疏，提高语料质量，从而改善最大熵模型的判别能力，另一方面，将针对识别错误的句子优化边界规则和特征模板选择；其次，尝试解决解决藏语复句以及嵌套语句的边界识别问题；最后，我们希望利用其他机器学习方法尝试解决藏语句子边界识别的问题，从而比较不同方法的优劣。

## 参考文献

[1]　赵维纳，刘汇丹，于新，等. 基于法律文本的藏语句子边界识别[C]//第五届全国青年计算语言学研讨会论文集，2010：480-486.

[2]　胡书津. 简明藏文文法[M]. 昆明：云南民族出版社，1988.

[3]　格桑居冕，格桑央京. 实用藏文文法教程[M]. 成都：四川民族出版社，2004.

[4]　扎西加，顿珠次仁. 自然语言处理用藏语格助词的语法信息研究[J]. 中文信息学报，2010，24(5)：41-45.

[5]　Riley, M. D. Some applications of tree-based modeling to speech and language indexing [C]//Proceedings of the DARPA Speech and Natural Language Workshop, 1989: 339-352.

[6]　Palmer, D. D., Hearst M. A. Adaptive Multilingual Sentence Boundary Disambiguation [J]. Computational Linguistics, 1997, 23(2): 241-269.

[7]　Liu, Y., Stolcke, A., Shriberg, E. and Harper, M. Using Conditional Random Fields for Sentence Boundary Detection in Speech[C]//Proc. ACL, 2005: 451-458.

[8]　Darroeh J. N. and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models [J]. The Annals of Mathematical Statistics, 1972, 43(5): 1470-1480.

[9]　胡书津. 书面藏语常用关联词语用法举要[M]. 昆明：云南民族出版社，1993.

[10]　格桑居冕. 藏语复句的句式[J]. 中国藏学，1996，(1)：132-141.

[11]　王诗文. 汉、藏语句子结构对比研究[J]. 西南民族大学学报(人文社科版)，2007，28(4)：50-55.

[12]　祁坤钰. 信息处理用藏文自动分词研究[J]. 西北民族大学学报(哲学社会科学版)，2006，(4)：92-97.

[13] 艾山·吾买尔,吐尔根·依步拉音. 统计与规则相结合的维吾尔语句子边界识别[J]. 计算机工程与应用, 2010,46(14): 162-165.

[14] Berger A. L., Della Pietra, S. A. and Della Pietra V. J. A Maximum Entropy Approach to Natural language Processing [J]. Computational Linguistics, 1996, 22(1): 39-71.

[15] Ratnaparkhi, Adwait. A Maximum Entropy Model for Part-of-speech Tagging [C]//Proceedings of EMNPL, 1996: 133-142.

[16] Mikheev, A. Tagging Sentence Boundaries [C]// Proceedings of ANLP-NAACL 2000: 264-271.

[17] Reynar, J. C. and Ratnaparkhi, A. A Maximum Entropy Approach to Identifying Sentence Boundaries [C]//Proceedings of the Firth Conference on ANLP, 1997: 803-806.

[18] Glenn Slayden, Mei-Yuh Hwang, and Lee Schwartz. 2010. Thai Sentence-Breaking for Large-Scale SMT [C]//Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing at COLING 2010: 8-16.

[19] Jiang Di. Identification of Boundaries of Object Clauses in Causative Verb Sentences in Modern Tibetan [J]. Journal of Chinese Language and Computing, 2006, 15(4): 185-192.

(上接第 15 页)

WG2 N2924R)[EB/OL]. [2005-02-08]. http://anubis.dkuug.dk/jtcl/sc2/wg2.

[2] Michael Evenson. Summary of repertoire for FDAM 5 0f ISO IEC 10646: 2003 (ISO/IEC JTC1/SC2/WG2 N3465) [EB/OL]. [2008-04-24]. http://anubis.dkuug.dk/jtcl/sc2/wg2.

[3] Michael Everson, Martin Hosken. Proposal for Encoding the Lanna Script in the BMP of the UCS(ISO/IEC JTC1/SC2/WG2 N3121R) [EB/OL]. [2006-09-09]. http://anubis.dkuug.dk/jtcl/sc2/wg2.

[4] 新华通讯社,等. GB/T 20092-2006,中文新闻信息置标语言[S]. 北京: 中国标准出版社,2006.

[5] 新华通讯社,等. GB/T 20093-2006,中文新闻信息分类与代码[S]. 北京: 中国标准出版社,2006.