# 添加冒号和分号分类标签特征的汉语逗号分类*

李艳翠[1,2,3]，谷晶晶[1,3]，周国栋[1,3]

（1. 苏州大学计算机科学与技术学院，江苏 苏州 215006；2. 河南科技学院信息工程学院，河南 新乡 453003;
3.苏州大学自然语言处理实验室，江苏 苏州 215006）

**摘要**：标点分析在句子和篇章分析中有重要作用，其中逗号的功能分类是标点分析的重点和难点。本文研究添加冒号和分号分类标签为特征的逗号自动分类。首先给出逗号、冒号和分号的分类方法，然后介绍基于此分类方法的逗号、冒号和分号标点分类语料库，最后分别考察添加冒号类别标签、分号类别标签以及同时添加冒号和分号类别标签为特征的逗号分类结果。实验结果表明，三种情况下的逗号分类正确率均有不同程度的提高。

**关键词**：逗号分类；冒号标签；分号标签；篇章分析；

**中图分类号**：TP391　　**文献标识码**：A

# Adding Colon and Semicolon Classification Label Feature to Chinese Comma Classification

LI Yancui[1, 2, 3], GU Jingjing[1, 3], ZHOU Guodong[1, 3]

(1. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China;
2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang , Henan 453003, China; 3.Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu 215006, China)

**Abstract:** Punctuation analysis plays an important role in the sentence and discourse analysis, in which the functional classification of the comma is the key and difficult point. This paper explores Chinese comma automatic classification based on adding the classification labels of Chinese colon or semicolon as new features. First, we describe the classification method of the comma, colon and semi-colon. Then introduce the corpus of comma, colon and semicolon based on this method. Finally, we investigate comma classification results by adding Chinese colon, semicolon and both two punctuations classification labels as new feature respectively. Experimental results show that the accuracy of comma classification improves in all three cases.

**Key words:** Chinese comma classification；Colon Labels；Semicolon Labels；Discourse Analysis；

## 1 引言

标点符号是书面语言的重要组成部分，同一种标点往往有不同的句法或篇章功能，如逗号有分隔小句、主谓关系和短语并列等不同的语言功能[1]，有效识别标点的功能，有助于句法分析、篇章分析、机器翻译等自然语言处理技术效果的提高。

在句法分析方面，李辛等[2]引入标点处理进行汉语长句句法分析，利用部分标点符号的特殊功能将复杂长句分割成子句序列，把整句的句法分析分成两级来进行，从而提高了复杂长句分析的正确率和召回率；Jin 等[3]提出利用逗号对汉语长句进行划分，通过汉语句子的上下文识别逗号左右两边的子句是并列关系还是从属关系，并利用这两种关系对逗号进行分类，进而提高句法分析的性能。在篇章分析方面，Xue 等[4]进行表示句子边界的逗号识别研究，提出逗号可等同于句子边界时要满足两点要求：一是逗号前后子句有完整的句法结构（即具有一个完整的 IP 结构，存在主谓宾），二是具有独立的句义且逗号前后子句间没有紧密的句法关系。Yang 等[5]对逗号的使用方

*收稿日期：2014-6-15 定稿日期：2014-7-28

基金项目：国家 863 计划前沿技术研究类项目（2012AA011102）；NSFC 面上项目(61273320)

作者简介：李艳翠(1982－ )，女，博士研究生，讲师，主要研究方向为自然语言处理；谷晶晶(1986－ )，女，硕士研究生，主要研究领域为自然语言处理；周国栋(1967－ )，男，博士，教授，博士生导师，通讯作者，主要研究领域为自然语言处理、多语言跨文本信息抽取。

法进行了更详细的分类，共分为七类：SB、IP_COORD、VP_COORD、ADJ、COMP、SBJ 和 Other。Yang 等采用了两种基于句法信息的方法实现逗号的自动分类。谷晶晶等[6]提出一种基于汉语句子的分词与词性标注信息做逗号自动分类的方法，结果表明利用词与词性进行逗号分类的方法是可行的。在机器翻译方面，黄河燕等[7]利用标点符号和关联词等把复杂长句进行切分，简化为多个独立的简单句，再进行翻译处理，以此提高机器翻译的性能。

从以上的研究可以发现，逗号功能识别是标点研究中的重点和难点，本文主要研究汉语逗号的功能分类。文献[8]统计显示汉语宾州树库(CTB6.0)中句号、问号、叹号、分号、逗号和冒号等标点的使用频率，其中句末点号句号、问号、叹号共占 29.55%，逗号高达 67.17%，其次是冒号（1.69%）和分号（1.85%）。由于逗号所占比例较大并且具有较多不同的功能，因此非常有必要进行逗号的功能分类研究。汉语句子中使用频率最高的除了逗号，还有冒号和分号，本文分别将 CTB6.0 语料中含有冒号和分号的句子抽取出来，进行逗号的自动分类识别实验。实验结果发现（见表 1）含冒号句子的语料和分号句子的语料中逗号自动分类的总体正确率上都严重低于全体语料的总体正确率，尤其是句子边界（SB）分类逗号的 F 值严重下降。说明含有冒号或分号的句子中逗号多元分类的自动识别效果不好，文献[6]中的错误分析中也指出了 IP_COORD 类与 SB 分类容易混淆。

表 1 全体语料与局部语料总体正确率对比

| 分类 | 指标 | 全体语料基准系统 | 含冒号语料基准系统 | 含分号语料基准系统 |
|---|---|---|---|---|
| All | Acc | 66.3 | 52.8 | 57.6 |
| SB | F | 63.2 | 37.3 | 30.0 |

说明：实验采用文献[6]的特征和最大熵分类器。含冒号语料是指从全体语料中抽取出来每个句子中至少包含一个冒号的语料；含分号语料是指从全体语料中抽取出来的每个句子中至少包含一个分号的语料。

逗号、冒号和分号在使用上存在一定的层次关系。通常情况下，分号的层次比逗号更接近根节点。在冒号作用域内，分号层次低于冒号，高于逗号。这些标点符号丰富的使用方法导致了汉语句子长度较长且语义复杂。逗号分类是标点分析的一个重要工作，由表 1 可知，含有冒号和分号的语料中逗号的分类效果较差，所以有必要专门进行处理，看能否增加逗号分类的正确率。

本文主要研究添加冒号和分号分类标签为特征后的逗号自动分类。主要从以下 3 方面进行展开：首先给出标点分类方法；然后介绍基于此分类方法的标点分类语料库；最后给出冒号和分号对逗号分类影响的实验结果与分析。

## 2 标点分类

### 2.1 逗号分类

本文是借鉴 Yang 等[5]提出的逗号分类标准，将逗号使用方法划分为 7 类。首先把逗号的使用方法在总体上分为连接的两子句之间存在关系和不存在关系。两子句之间存在的关系又分为并列关系和从属关系。并列关系有 3 种类型（SB、IP_COORD 与 VP_COORD），从属关系也有 3 种类型（ADJ、COMP 与 SBJ）。每种类别的具体说明见文献[6]，图 1 展示了逗号分类类别。下面对每种类别进行简单说明，实例中属于此类的逗号用 c1…cn 标识，如例 1 中的 c1 和 c2 属于类别 SB，例 2 中的 c3 属于 IP_COORD 类。

**SB(Sentence Boundary)**：分割句子边界的逗号。该类逗号是指在某些语境下，起句子边界的作用。该类逗号要求逗号左右的子句都是 IP 结构，父节点为根节点。如例 1 中的 c1 和 c2。

例1: 陕西省目前批准的外资项目已达两千四百多个，c1 协议利用外资额四十多亿美元，c2 实际引进外资超出十六亿美元。

**IP_COORD (IP Coordination)**：分割父节点为非根节点的并列 IP 结构的逗号。如 c3 和 c4。

例2: 周永康在会议工作报告中指出，陆上石油勘探开发遇到一系列世界级难题，c3 投资成本日益上升，c4 企业改革和产业结构调整任务艰巨。

**VP_COORD (VP Coordination)**：分割并列动宾短语的逗号。这一类的逗号与 IP_COORD 类逗号相似，都是分割嵌套结构中的并列结构。

例3:　中国银行是四大国有商业银行之一，c5 也是中国主要的外汇银行。

**ADJ (Adjunction)**：分割附属从句与主句的逗号。附属从句是指在句子中担当某种句子成分的主属结构。虽然从句部分的句子结构是完整的，但它并不能脱离主句部分独立完整的表达意思。

例4:　为了在运行机制上与保护区相配套，c6 宁波保护区率先在中国实施了企业依法注册直接登记制的试行一站式管理。

**COMP (Complementation)**：分割句子谓语与宾语的逗号。通常出现在“表示”、“指出”、“认为”、“介绍”等提示性动词之后。

例5:　西谚说，c7 知识就是力量。

**SBJ (Sentential Subject)**：分割句子主语和谓语的逗号。SBJ 类逗号表示的是逗号分割开了句子的主语与动宾结构。

例6:　出口快速增长，c8 成为推动经济增长的重要力量。

**Other**：其他类型。本文将不属于上述 6 种类型的逗号都划分为 Other 类型。

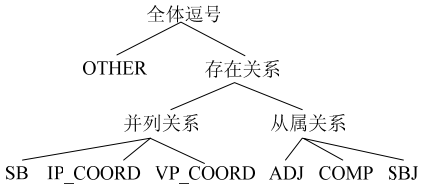


图 1 逗号分类类别

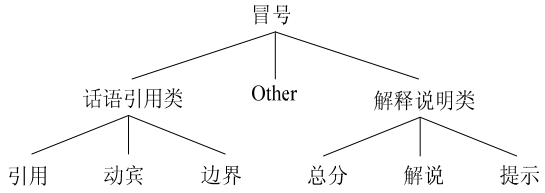


图 2 冒号分类标准

## 2.2 冒号分类

参考文献[1]，本文将冒号的使用方法归纳为 7 类（如图 2）：引用、动宾、边界、总分、解说、提示、Other。其中引用、动宾和边界又归为话语引用类，而总分、长解说和短解说又归为解释说明类。Other 分类是对冒号的一些不经常使用的用法归类。下面对每种类别的冒号进行举例说明。

**引用(Nm)**：指人物专名或人物代指与该人所说的话被冒号分开的情况。该类冒号，通常出现在侧重对话内容的记录中，这种对话一般不关注对话人的语气、表情、动作等。

例7:　秦牧：c9 要学好语文，必须注意多读、多写、多思索。

**动宾(VP)**：该类冒号分割开了谓语动词与宾语。常用的谓语动词有：问、答、说、曰、云、想、是、证明、宣布、例如、如下等。

例8:　克莱因说：c10“普遍的观点是人以群分，人们总喜欢和自己相似的人，所以有理论提出多样化不利于团结。”

**边界(SB)**：该类冒号被定义为句子边界，冒号前后的句子都是一个完整的 IP 结构，可独立存在。冒号后的句子一般是对冒号前句中主语的话语引用，由左右双引号界定。

例9:　凤姐连忙告诉小丫头传饭：c11“我和太太都跟着老太太吃。”

**总分(ZF)**：冒号前的句子是总说，冒号后面的句子是对前面句子的分说。

例10:　本文将冒号的使用方法归纳为七类：c12 引用、动宾、边界、总分、短解说、提示、Other。

**解说(LJ)**：后面的句子是对冒号前面的词语的解释说明。

例11:　有人曾做过对比实验：c13 两个病情相近，年龄和体重相差无几的手术患者，每天食用一只海参的患者，会比另一个患者提前 20 天左右全面康复。

**提示(SJ)**：该类是生活中常用的、位于提示短语后的冒号。该类冒号是从解说类中分离出来的一类，冒号后的内容也是对冒号前词或短语的解说，该类冒号前通常只有一个词或短语。

例12:　电话：c14 8888888

**Other**：本文设置一个 Other 类，是因为存在一些使用方法出现频率较低的冒号，有分总类冒号、呼语类冒号以及作者与作品之间的冒号，例如“朱自清：《背影》”。这些使用方法的冒号都可单独作为一类，但由于实际语料中出现的频率较低，故将这些使用方法统归为 Other 类。

### 2.3 分号分类

参考文献[1]，本文对分号设置 3 类标注标签，分别是：并列关系(BL)、非并列关系(FB)和条款(TK)。其中，并列关系是指分号两边的多个子句是并列的关系，而非并列关系是指两边的多个子句间存在转折、因果等非并列关系。条款类是指分条或分行列举的分句之间使用分号，这类冒号通常用在冒号的作用域内。标注方法与标注冒号的分类标签方法相同。

例13: 语言，人们用来抒情达意；c15 文字，人们用来记言记事。

例14: 我国年满十八周岁的公民，不分民族、种族、性别、职业、家庭出身、宗教信仰、教育程度、财产状况、居住年限，都有选举权和被选举权；c16 但是依照法律被剥夺政治权力的人除外。

例15: 中华人民共和国行政区域划分如下：c17（一）全国分为省、自治区、直辖市；c18（二）省、自治区分自治州、县、自治县、市；c19（三）县、自治县分乡、民族乡、镇。

例 13 中的分号为并列关系类，例 14 中的分号属于非并列关系类，例 15 中的分号属于条款类。对于条款类的分号，有时一个分句为一行，如例 15 中的（一）（二）（三）可以分别作为一个段落，这时的分号相当于段落间的分割符号。识别该类分号对于基于段落的篇章分析有一定的帮助。

## 3 标点分类语料

### 3.1 逗号分类语料

据统计，CTB 6.0 语料中共有 51886 个逗号，各分类所占的逗号数量比例如表 2 所示。采用与文献[6]中相同的训练语料和测试语料划分方式，训练语料包含了 42497 个逗号，测试语料包含了 5436 个逗号。

表 2 CTB 6.0 语料中各类逗号分布

| 类型 | 数量 | 所占百分比(%) | 类型 | 数量 | 所占百分比(%) |
|---|---|---|---|---|---|
| SB | 11062 | 23 | COMP | 3053 | 6.4 |
| IP_COORD | 3861 | 8 | SBJ | 2165 | 4.5 |
| VP_COORD | 9445 | 19.8 | Other | 14677 | 30.6 |
| ADJ | 3689 | 7.7 | 所有逗号 | 47952 | 100 |

### 3.2 冒号分类语料

本文的冒号语料实验数据，是从逗号自动分类与识别语料（CTB6.0）中抽取出来的。抽取出的冒号语料大小为原始全体语料的 9%，具体标注的冒号数量和冒号语料中逗号的数量如下表 3 所示。由表 3 可以看出，语料中含有的冒号的个数只是逗号个数的 50%左右，但是位于冒号后的逗号占逗号总数的 78%。由此也可以预见，添加冒号分类标签特征后，将对逗号的自动分类与识别产生影响。在逗号分类的训练语料和测试语料中分别抽出所有包含冒号的句子，构成新的训练语料和测试语料。对抽取出来的训练语料和测试语料，首先分别进行预处理，再分别进行人工标注汉语冒号分类标签。所标注的冒号分类标签参考 2.2 中的冒号分类，主要标注 7 类标签，分别是引用（Nm）、动宾（VP）、边界（SB）、总分（ZF）、解说（LJ）、提示（SJ）和 Other。

冒号语料中存在与例 16 类似的句子，即句子中只含有冒号而没有逗号，且冒号位于句末，这种情况的句子不在本文实验的考察范围之内。类似例 16 中的冒号一般是位于一个段落的结尾处，下面紧跟着的一个段落或者是多个段落都在该冒号作用域内，但这些段落中的逗号分类与识别已经不受该冒号的影响，故该类冒号不在本文的考察范围之内。

例16: 港台会师看新局:

表 3 冒号语料中各标点个数

| 类别 | 训练语料 | 测试语料 |
|---|---|---|
| 冒号个数 | 1756 | 225 |
| 逗号个数 | 3287 | 443 |

表 4 分号语料中各标点个数

| 类别 | 训练语料 | 测试语料 |
|---|---|---|
| 分号个数 | 765 | 89 |
| 逗号个数 | 2365 | 288 |

### 3.3 分号语料

分号语料的实验数据，同样是从逗号自动分类与识别语料中抽取出来的。采取和冒号语料同样的处理方法，经过预处理后再进行人工标注。

对分号语料中含有的分号和逗号个数进行统计结果如表 4 所示。据统计，抽取出的分号语料大小为原始全体语料的 5.5%。相比于冒号，分号数量更少。

## 4.实验结果与分析

本节分别进行了添加冒号分类标签特征、添加分号分类标签特征和同时添加这两种标点分类标签特征的实验。这 3 个实验采用了基本相同的方法，流程如图 3 所示。根据 Yang 等人[5]一文中介绍的逗号各分类对应的句法模型，预处理系统每次读入一个带句法信息的句子，从句中逗号，分别提取逗号分类的三元组文件，即[句子标号，逗号序号，逗号分类标签]。通过对 CTB 6.0 句法树库的自动提取（即预处理系统），可以得到该实验训练模型时所需要的逗号训练样例（即三元组文件）和测试样例。

本文基本特征选取和文献[6]相同：1) 子句主干特征，从分词与词性标注的序列中，选取 3 个能表示子句主干的词。2) 当前逗号序号及序号前的逗号分类类别，通过提取这些特征可以间接反映句子的层次结构。3) 词汇特征，提取词汇特征是为了得到体现逗号左右子句特点的词，比如存在介词、连词、副词等。另外，分别添加冒号或分号的分类标签为一组新特征。

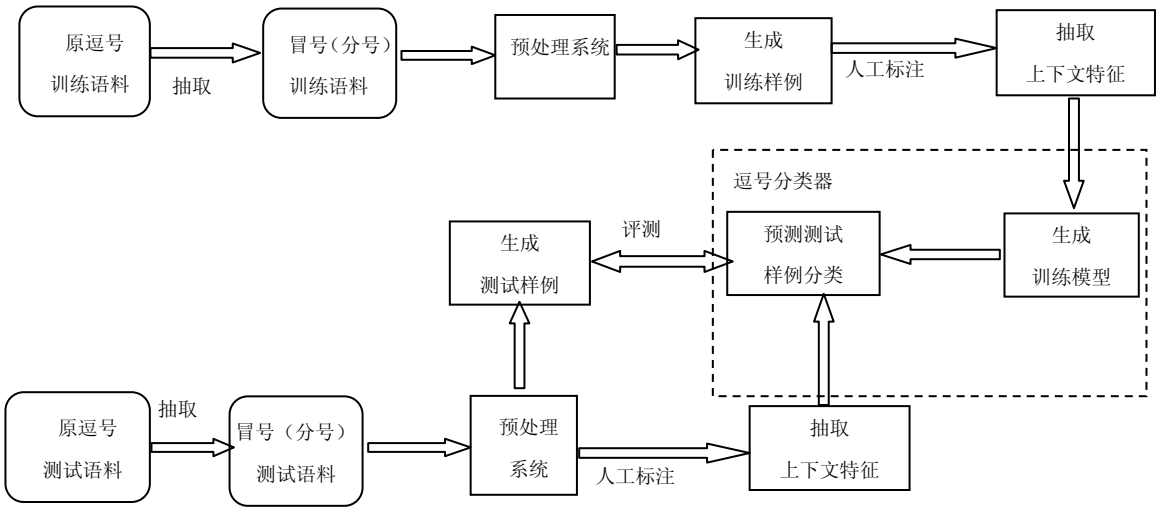


图 3 添加冒号（分号）分类标签特征的逗号分类流程图

### 4.1 添加冒号分类标签特征的实验结果及分析

#### 4.1.1 冒号语料的实验结果

按照文献[6]的最大熵模型实验提取上下文特征的方法，在提取原特征的基础上，将当前逗号前的冒号分类标签作为一个新的特征加入到特征集合中。实验的结果如表 5 所示。

从表 5 可以看出，逗号分类的自动识别整体正确率提高了 9.9%，说明通过添加冒号分类标签特征来提高逗号自动识别正确率的方法是可行的，而这两类标点符号之间是存在影响的。表 5 中，各分类逗号的 F 值都有不同程度的提高，尤其是 SB 分类和 IP_COORD 分类，分别提高了 32.3% 和 23.0%。说明添加的冒号分类标签，对这两类逗号识别正确率影响最大，一些被错分为 SB 分类

的逗号，在本实验中被正确识别为 IP_COORD 分类。至于 SBJ 分类的自动识别 F 值为零，是由于属于该分类的逗号在训练样例中只出现了 3 次，在测试样例只有 1 个。

表 5 冒号语料中逗号自动识别结果

| 分类 | 指标 | 冒号语料基准系统 | 添加冒号分类标签特征 |
|---|---|---|---|
| All | Acc | 52.8 | **62.7** |
| SB | P R F | 36.8 37.8 37.3 | 65.5 74.3 **69.6** |
| IP_COORD | P R F | 45.7 38.9 42.0 | 59.3 72.0 **65.0** |
| VP_COORD | P R F | 55.8 77.4 64.9 | 68.3 78.2 72.9 |
| ADJ | P R F | 46.2 22.2 30.0 | 54.5 22.2 31.6 |
| Comp | P R F | 62.5 45.5 52.6 | 66.7 54.5 60.0 |
| SBJ | P R F | 0 0 0 | 0 0 0 |
| Other | P R F | 63.7 61.4 52.6 | 66.9 63.6 65.2 |

**4.1.2 全体语料的实验结果**

在冒号语料的实验取得成功后，本实验将标注了冒号分类标签的语料带入到全体语料中，替换没有被标注的冒号句子。在标注了冒号分类标签的全体语料上，再次进行实验，新实验同样是在添加冒号分类标签特征后进行多元逗号分类。实验结果如表 6 所示。

表 6 添加冒号分类标签特征的全体语料实验结果对比

| 分类 | 指标 | 全体语料基准系统 | | 添加冒号分类标签特征 | |
|---|---|---|---|---|---|
| | | Maxent | CRF | Maxent | CRF |
| All | Acc | 66.3 | 68.1 | **67.0** | **68.9** |
| SB | P R F | 57.5 70.2 63.2 | 60.8 72.3 66.1 | 59.8 70.2 **64.6** | 61.9 72.8 **66.9** |
| IP_COORD | P R F | 55.3 27.9 37.1 | 62.1 38.5 47.6 | 56.0 34.4 **42.6** | 65.4 43.7 **52.4** |
| VP_COORD | P R F | 63.8 75.6 69.2 | 64.0 77.1 70.0 | 63.7 75.3 69.0 | 65.3 77.6 70.9 |
| ADJ | P R F | 61.2 50.1 55.1 | 62.0 51.7 56.4 | 61.5 52.0 56.4 | 62.1 52.0 56.6 |
| Comp | P R F | 89.6 90.8 90.2 | 88.2 91.8 90.0 | 88.7 91.2 90.0 | 87.9 91.5 89.7 |
| SBJ | P R F | 53.3 12.5 20.3 | 46.2 9.4 15.6 | 50.0 10 15.8 | 46.7 10.9 17.7 |
| Other | P R F | 75.8 69.6 72.5 | 78.0 69.1 73.3 | 76.1 69.9 72.8 | 77.7 69.3 73.3 |

表 6 列出了添加冒号分类标签前后，分别采用最大熵模型和 CRF 模型的实验结果。基于最大熵模型的全体语料整体正确率提高了 0.7%，基于 CRF 模型的全体正确率提高了 0.8%，由此也再次说明基于 CRF 模型的自动分类识别正确率要高于基于最大熵模型的自动识别正确率。由表 3 统计的数据可知，冒号语料中的逗号个数占全体语料中逗号个数的 6.9%，而由表 5 添加冒号分类标签特征的冒号语料逗号分类总体正确率提高 9.9%，表 6 全体语料总体正确率提高 0.8%，实验说明冒号语料和全体语料在添加冒号分类标签特征后，提高的总体正确率是成比例的。

同时，SB 分类和 IP_COORD 分类的逗号在全体语料的实验中，结果都有一定的提高。在全体语料上，SB 分类并没有 IP_COORD 分类 F 值提高的多，因为在全体语料中，SB 分类共有 1311 个，而 IP_COORD 分类只有 506 个。

**4.1.3 边界识别**

引言中提到冒号对 IP_COORD 分类和 SB 分类的逗号存在明显影响，由于 SB 分类属于逗号标示句子边界的情况，所以本文将同样考察冒号对识别逗号作为句子边界情况存在的影响。识别 SB 分类，即为识别句子边界(EOS，End Of a Sentence)。结合本文的实验，只需将 SB 分类归为 EOS，余下的 6 类归为非句子边界(Non-EOS，Not the End Of a Sentence)。表 7 列出了基于最大熵模型的

全体语料在添加冒号标签特征前后，识别逗号标示句子边界的实验结果。

表 7 逗号标示句子边界的识别结果

| 类别 | 基准系统 | | | 冒号标签系统 | | |
|---|---|---|---|---|---|---|
| 指标 | P | R | F | P | R | F |
| Overall | - | - | 80.3 | - | - | **81.5** |
| EOS | 57.5 | 70.0 | 63.2 | 59.8 | 70.2 | 64.6 |
| NEOS | 89.8 | 83.5 | 86.5 | 90.0 | 85.1 | 87.5 |

由表 7 可以看出，在添加冒号标签特征后，逗号标示句子边界的实验结果在总体正确率上提高 1.2%，EOS 和 NEOS 分类的 F 值也分别有所提高。再次说明，冒号分类标签对逗号的分类自动识别存在影响。

### 4.2 添加分号分类标签特征的实验及分析

#### 4.2.1 分号语料的实验结果

添加分号分类标签特征的实验与添加冒号分类标签特征的实验类似。在提取原有特征的基础上，将当前逗号前的分号分类标签作为一组新的特征添加到特征集合中。实验的结果如表 8 所示。

表 8 中分号语料基准系统的实验是基于最大熵模型的，添加分号分类标签特征的实验分别采用了最大熵和 CRF 两种模型。CRF 模型的自动识别正确率比最大熵模型的更高，但这里主要对比添加分号分类标签特征前后的最大熵模型的实验结果。由表 8 可知，基于最大熵模型的实验结果中，逗号分类的自动识别整体正确率提高了 4.6%。

表 8 分号语料中逗号分类自动识别结果及对比

| 分类 | 指标 | 分号语料基准系统 | | | 添加分号分类标签特征 Maxent | | | 添加分号分类标签特征 CRF | | |
|---|---|---|---|---|---|---|---|---|---|---|
| All | Acc | 57.6 | | | **62.2** | | | **64.9** | | |
| SB | P R F | 27.8 | 32.6 | 30.0 | 32.0 | 34.8 | 33.3 | 40.8 | 43.5 | 42.1 |
| IP_COORD | P R F | 53.3 | 20.0 | 29.1 | 61.5 | 20.0 | 30.2 | 77.3 | 42.5 | 54.8 |
| VP_COORD | P R F | 59.8 | 82.7 | 69.4 | 64.9 | 91.4 | 75.9 | 63.5 | 90.0 | 74.5 |
| ADJ | P R F | 57.1 | 26.7 | 36.4 | 55.6 | 33.3 | 41.7 | 60.0 | 40.0 | 48.0 |
| Comp | P R F | 87.5 | 87.5 | 87.5 | 87.5 | 87.5 | 87.5 | 100.0 | 87.5 | 93.3 |
| SBJ | P R F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Other | P R F | 70.1 | 67.7 | 69.1 | 73.4 | 71.9 | 72.6 | 75.3 | 66.7 | 70.7 |

表 8 中，各分类逗号的 F 值都有不同程度的提高，但并不像添加冒号分类标签的实验结果中 SB 分类和 IP_COORD 分类正确率提高的幅度那样大。正确率提高相对较高的是 ADJ 类逗号和 VP_COORD 类逗号。实验表明添加分号分类标签特征提高逗号自动识别正确率的方法是可行的。

#### 4.2.2 全体语料的实验结果

在分号语料的实验取得成功后，本文同样将已标注的分号语料反馈到原语料中。同样的方法，实验结果如表 9 所示。

由表 9 可知，添加新特征后最大熵模型的总体正确率提高了 0.2%，而 CRF 模型的总体正确率提高了 0.5%。在添加冒号分类标签特征的实验结果表 6 中，CRF 模型和最大熵模型分别提高了 0.7%和 0.8%。效果不明显与冒号语料所占的比例有关，由 3.2 和 3.3 可知，冒号语料占全体语料的 9%，而分号语料明显较小，占全体语料的 5.5%。

表 9 添加分号标签后的全体语料实验结果及对比

| 分类 | 指标 | 全体语料基准系统 | | 添加分号分类标签特征 | |
|---|---|---|---|---|---|
| | | Maxent | CRF | Maxent | CRF |
| All | Acc | 66.3 | 68.1 | **66.5** | **68.6** |
| SB | P R F | 57.5 70.2 63.2 | 60.8 72.3 66.1 | 58.1 68.6 63.0 | 61.3 71.9 66.2 |
| IP_COORD | P R F | 55.3 27.9 37.1 | 62.1 38.5 47.6 | 50.9 27.5 35.7 | 63.2 40.1 49.1 |
| VP_COORD | P R F | 63.8 75.6 69.2 | 64.0 77.1 70.0 | 63.9 76.1 69.5 | 65.1 78.3 71.1 |
| ADJ | P R F | 61.2 50.1 55.1 | 62.0 51.7 56.4 | 60.8 50.9 55.4 | 61.9 52.3 56.5 |
| Comp | P R F | 89.6 90.8 90.2 | 88.2 91.8 90.0 | 88.5 91.8 90.2 | 88.3 92.5 90.4 |
| SBJ | P R F | 53.3 12.5 20.3 | 46.2 9.4 15.6 | 47.1 12.5 20.0 | 38.9 10.9 17.1 |
| Other | P R F | 75.8 69.6 72.5 | 78.0 69.1 73.3 | 76.7 70.9 73.7 | 78.1 69.7 73.7 |

比较表 6 和表 9 可知，CRF 模型比最大熵模型效果要好。因为 CRF 模型计算了全局最优的输出节点的条件概率，而不是只通过当前的状态来定义下一个节点的状态。通过分析冒号和分号的作用域可以发现，冒号的作用域是从冒号后的第一个字符开始到句末标点结束；而分号的作用域不止包含在分号后面的句子部分，它的作用域为当前分号前后相邻的两个分号（相邻不是分号时，为句子开始字符和句子结束字符）之间。故在添加分号分类标签特征的实验中，更能体现 CRF 模型的优越性。

### 4.3 同时添加冒号和分号分类标签特征的实验

同时添加冒号和分号分类标签为特征的实验，是指同时添加当前逗号前的冒号的分类标签和分号的分类标签作为一组新的特征进行实验。实验结果如表 10 所示。

表 10 添加分号和冒号两类标签后的全体语料实验结果及对比

| 分类 | 指标 | 全体语料基准系统 | | 添加分号和冒号分类标签特征 | |
|---|---|---|---|---|---|
| | | Maxent | CRF | Maxent | CRF |
| All | Acc | 66.3 | 68.1 | **67.2** | **69.2** |
| SB | P R F | 57.5 70.2 63.2 | 60.8 72.3 66.1 | 59.8 70.8 64.8 | 62.4 73.4 67.5 |
| IP_COORD | P R F | 55.3 27.9 37.1 | 62.1 38.5 47.6 | 55.5 34.0 42.2 | 65.5 44.3 52.8 |
| VP_COORD | P R F | 63.8 75.6 69.2 | 64.0 77.1 70.0 | 63.9 74.5 68.7 | 65.5 77.8 71.1 |
| ADJ | P R F | 61.2 50.1 55.1 | 62.0 51.7 56.4 | 61.9 52.0 56.5 | 61.0 51.7 56.0 |
| Comp | P R F | 89.6 90.8 90.2 | 88.2 91.8 90.0 | 89.9 90.8 90.4 | 88.3 91.8 90.0 |
| SBJ | P R F | 53.3 12.5 20.3 | 46.2 9.4 15.6 | 53.3 12.5 20.5 | 43.8 10.9 17.5 |
| Other | P R F | 75.8 69.6 72.5 | 78.0 69.1 73.3 | 76.3 70.4 73.3 | 78.3 69.5 73.6 |

通过对全体语料的基准系统和分别添加其中某一个标点的分类结果相比，该综合实验的总体正确率及各项的分类的 F 值都有所提高，说明本文提出的添加其他标点符号的分类标签特征辅助逗号多元分类的自动识别方法是可行的，且取得了相对较好的成绩。CRF 模型的总体正确率达到 69.2%，已经非常接近 Yang 等基于句法信息的 71.5%的总体正确率。

## 5 结论

本文主要研究了分别添加冒号和分号分类标签，以及同时添加两类标点的分类标签特征后，对逗号自动分类结果的影响。实验结果表明，在分别添加冒号或分号分类标签特征后，逗号多元分类的自动识别正确率都有所提高，且在同时添加这两类标点分类标签特征时，逗号识别的正确率达到 69.2%。本文实验说明分号和冒号分类对逗号分类是存在影响的，合理的利用冒号或分号分类标签可以提高逗号分类的正确率。

## 参考文献


[1] 中华人民共和国国家质量监督检验检疫总局、中国国家标准化管理委员会. GB/T15834-2011 标点符号用法[M]. 北京:中国标准出版社, 2011

[2] 李幸, 宗成庆. 引入标点处理的层次化汉语长句句法分析方法[J]. 中文信息学报, 2006, 20(4): 8–15

[3] Jin M X, Kim M Y, Kim D, et al. Segmentation of Chinese long sentences using commas[C].// Proceedings of 3rd ACL SIGHAN Workshop. Barcelona,2004:1-8

[4] Nianwen Xue, Yaqin Yang. Chinese sentence segmentation as comma classification. [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011

[5] Yaqin Yang and Nianwen Xue. Chinese Comma Disambiguation for Discourse Analysis. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics (ACL). 2012

[6] 谷晶晶, 周国栋. 基于分词与词性标注的汉语逗号自动分类[J]. 计算机工程与应用，2014.DOI:10.3778/j.issn.1002-8331.1310-0034.

[7] 黄河燕, 陈肇雄. 基于多策略分析的复杂长句翻译处理算法[J]. 中文信息学报, 2002, 16(3): 1–7

[8] 李艳翠, 冯文贺, 周国栋. 基于逗号的汉语子句识别研究[J].北京大学学报,2013,49(1):7-14