基于分词与词性标注的汉语逗号自动分类

谷晶晶，周国栋

(苏州大学计算机科学与技术学院，江苏，苏州，215006)

E-mail: 20114227022@suda.edu.cn

摘 要：近年来，标点符号作为篇章的重要部分逐渐引起研究者的关注。然而，针对汉语逗号的研究才刚刚展开，采用的方法也大多都是在句法分析的基础上，尚不存在利用汉语句子的表层信息开展逗号自动分类的研究。本文提出了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，并采用了两种有监督的机器学习分类器，即最大熵分类器和 CRF 分类器，来完成逗号的自动分类。在 CTB 6.0 语料上的实验表明，CRF 的总体结果比最大熵的要好，而这两种分类器的分类精度都非常接近基于句法分析方法的分类精度。由此说明，基于词与词性做逗号分类的方法是可行的。

关键词：汉语逗号分类；最大熵；CRF

Chinese Comma Classification Based on Segmentation and Part of Speech Tagging

Gu Jingjing ， Zhou Guodong

(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006)

E-mail:20114227022@suda.edu.cn

Abstract: In recent years, punctuation as an important part of discourse is attracting more and more attention of the researchers. However, most methods were based on syntactic analysis. Research of Chinese comma classification using the surface information of Chinese sentences does not exist. This paper proposes a method for Chinese comma classification based on segmentation and part-of-speech tagging and adopts two supervised machine learning classifiers, namely the maximum entropy classifier and CRF classifier, to complete the automatic classification of commas. Experimental results on the CTB 6.0 corpus show that CRF model is better than maximum entropy model, and the accuracy of the two classifiers are very close to the method based on syntactic analysis. It demonstrates that the method for Chinese comma classification based on segmentation and part-of-speech tagging is feasible.

Keyword: Chinese Comma Classification; Maximum Entropy; CRF

1. 引言

汉语中常用的标点符号有逗号、句号、问号等 12 种，其中逗号是最常见的句中停顿符号，它有二十多种不同的使用方法，如主谓之间停顿、谓语与宾语之间停顿、状语与主句之间停顿、流水句之间停顿等。本文将通过下面的三个例子，对逗号的具体使用方法做一个简单的考察。在本文中，把长句子被逗号分隔开的每个片段统称为子句。

(1) 有这个良好的开端，p1 人们完全有理由相信，p2 东亚国家间的协调工作在不久的将来必将展现更加美好的前景 。

(2) 陕西省目前批准的外资项目已达二千四百多个，p3 协议利用外资额四十多亿美元，p4 实际引进外资超过十六亿美元。

(3) 周永康在会议工作报告中指出，p5 陆上石油勘探开发遇到一系列世界级难题，p6 投资成本日益上升，p7 企业改革和产业结构调整任务艰巨。

例句(1)由三个子句构成，逗号 p1 的左子句与右子句是状语与主语之间的停顿。逗号 p2 的左右子句存在谓语与宾语之间的停顿。例句(2)则是很常见的流水句，该句中的逗号是流水句间的停顿，句中的三个子句是相互独立的，这里的逗号 p3 和逗号 p4 相当于句号。例句(3)中，逗号 p6 和逗号 p7 连接的三个子句也是相互独立的，但这三个子句是“指出”的

宾语，所以这两个逗号不能等同于句号。

目前，从自然语言处理角度对汉语逗号的自动分类研究开展的很少，现有的大多数自动句法分析系统也都忽略了标点符号的特殊作用。本文将从自然语言处理角度研究逗号的分类问题，并提出了一种基于句子的词与词性信息做逗号自动分类的方法，以期有助于汉语篇章自动分析。

本文采用了Yang and Xue (2012)[8]一文中提出的逗号分类标准。Yang and Xue的逗号自动分类方法是在句法分析的基础上，而我们发现仅利用句子的词与词性信息，对逗号做自动分类也可以取得较好的效果。我们选择了三组特征：a.子句主干特征，b.当前逗号序号及序号前的逗号分类类别特征，c.词汇特征，并使用了两种逗号自动分类器，分别是最大熵分类器和CRF分类器。

本文的组织结构如下：第 2 节简单介绍逗号的相关研究工作；第 3 节阐述逗号分类标准和逗号自动分类方法；第 4 节给出实验结果及实验分析；第 5 节给出结论，并对下一步工作进行展望。

2. 相关工作

近年来，随着对标点符号研究工作的展开，对逗号的研究也逐渐得到关注，比如 Jin 等 (2004)[3]提出利用逗号对汉语长句子进行划分。该文章主要识别逗号左右两边的子句是并列关系还是从属关系。

在 Li 等 (2005)[7]的文章中介绍了关于层次化汉语长句结构分析，提出了一种针对汉语长句子句法分析的分层处理方法。该方法用标点符号(包括逗号、分号和冒号)对长句子进行切分，然后对切分单元分别处理，得到各部分的分析子树，最后将子树合并，形成完整的句法分析树。该文揭示了基于标点符号的层次化汉语长句结构分析的优越性。

Xue and Yang(2011)[5]这篇文章中，主要研究了如何识别哪些逗号等同于句子边界的情况。并提出了逗号可等同于句子边界时要满足的两点要求：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。对应的句法树结构为，逗号左右两边的兄弟节点为 IP，父节点为根节点。如例句(2)中的逗号 p3 和逗号 p4 就是句子边界，而逗号 p6 和逗号 p7 则不是句子边界。本文同样对逗号等同于句子边界时进行了自动识别。

Yang and Xue (2012)[8]一文中，对逗号的使用方法进行了更详细的分类，共分为七类，其中也包含了逗号等同于句子边界的情况。Yang and Xue 采用了两种逗号自动分类方法：第一种方法是在加入逗号分类信息的情况下，运用句法分析器对句子重新做句法分析，由此得到逗号的分类；第二种方法是在句法分析器做完句法分析的基础上，建立最大熵逗号分类器做逗号自动分类。本文是采用 Yang and Xue (2012)[8]给出的逗号分类标准，提出了我们基于句子的分词与词性信息做逗号自动分类的方法，该方法是以前没有使用过的。

3. 逗号分类标准及逗号自动分类方法

3.1. 逗号分类标准

我们采用了 Yang and Xue (2012)[8]一文中的逗号分类标准，共将逗号共分为七类。首先，把逗号的使用方法总体上分为两大类，一类是逗号连接的两子句之间存在关系，另一类是两子句之间不存在关系。两子句之间存在的关系又可以分为并列关系和从属关系。并列关系又可分为三种类型(SB、IP_COORD 与 VP_COORD)，从属关系也分为三种类型(ADJ、COMP 与 SBJ)。其中，SB 分类属于逗号等同句子边界的情况；IP_COORD 分类属于逗号的左右子句有完整的 IP 结构但句法上却不独立的情况，上面例句中的逗号 p6 和逗号 p7 就属于该分

类；VP_COORD 分类属于并列的动宾短语，由共同主语发出的一系列动作；ADJ 分类属于从句与主语之间的停顿；COMP 分类属于谓语与宾语之间的停顿；SBJ 属于主语与谓语之间的停顿。图 1 给出了逗号分类类别。

根据 Yang and Xue (2012)[8]一文中介绍的逗号各分类对应的句法模型，通过对 CTB 6.0 句法标注语料的自动提取，可以分别得到该实验训练模型时所需要的逗号分类数据，和评测时所需要的逗号分类标准答案。

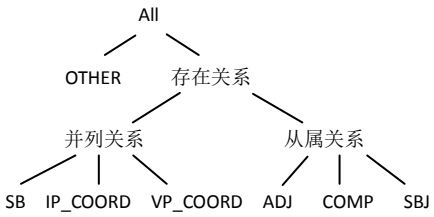

图 1 逗号分类类别

3. 2. 逗号自动分类方法

采用上一小节介绍的逗号分类标准，本文提出了一种基于句子的词与词性信息做逗号自动分类的方法。我们选择了三组特征，采用了最大熵和 CRF 两种有监督的机器学习模型。在实现逗号自动分类过程中，首先从分词与词性标注的句子中提取特征，分别标注训练语料集和测试语料集，再利用训练语料集训练出来一个最大熵(或 CRF)模型，最后根据已训练的最大熵(或 CRF)模型，对测试语料集进行逗号自动分类。如下面的图 2 展示了我们的逗号分类器。本文采用的最大熵分类器，是由张乐编写的 Maxent 工具包[1]；而 CRF 分类器，是运用 CRFsuite 工具包[2]。CRFsuite 可以像 Maxent 一样通过程序提取特征，只需加一些句子的开始和结束的标记即可，不需要写特征模板 template。CRFsuite 经常被用来处理列不整齐的文本，例如本文语料中的逗号左右子句就是不定长的。我们在构建不同的逗号分类器时，提取了相同的特征。下面将讲述如何提取特征。

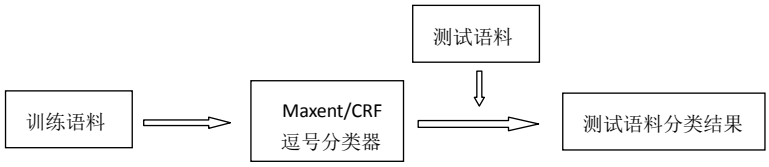

图 2 逗号分类器

3. 3. 特征提取

本文对每个逗号进行独立处理，主要对每个逗号的左右子句进行特征提取，但是每个逗号之间并不是绝对独立的，故在特征提取过程中加入了当前逗号前的句中逗号分类信息。本文共提取了三组特征：a.子句主干特征，从分词与词性标注的序列中，选取三个能表示子句主干的词。b.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。c.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。将通过下面的例句(4)，展示为逗号 p8 和逗号 p9 做自动分类所需要提取的特征，并在表 1 中列出。

(4) 中国_NR 银行_NN 是_VC 中国_NR 四_CD 大_JJ 国有_JJ 商业_NN 银行_NN 之一_NN ，_PU p8 也_AD 是_VC 中国_NR 的_DEG 主要_JJ 外汇_NN 银行_NN ，

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html
[2] http://www.chokkan.org/software/crfsuite/

_PU p9 它_PN 在_P 海内外_NN 拥有_VV 较为_AD 发达_VA 的_DEC 分支_NN 机构_NN 网络_NN 。_PU

3.3.1 子句主干特征

一个句子的主干由主谓宾构成。但由于本文的方法是基于分词与词性标注的句子，在没有句法分析的情况下，无法得知句中的主谓宾信息。所以，本文将选取三个具有主谓宾的词性特征、句法特征和修辞特征的词，分别称为词 a、词 b 和词 c，由此间接得到句子的主干信息。越能够准确的定位这三个词，越有助于提高逗号自动分类模型的正确率。因此，所有词的提取都要基于规则。在子句中先选取词 b，再选取词 a 和词 c，词 b 选取后需作位置标记。

词 b：选取词性为动词（即词性标记为 VV、VC、VA、VE 的词）或数词（CD）的这样一个词。依次选取规则：

(i) 优先选择词性标记为“VC” 的动词；
(ii) 其次选择出现“VV”+“AS”结构的动词，即谓语+补语形式；
(iii) 选择“AD”+“VV”结构的动词，即有副词修饰的动词，属于状语修饰谓语的情况；
(iv) 选择“把字句”中的标记为“BA”的词；
(v) 选择子句中普通动词词性标记的最后一个动词
(vi) 最后在没有任何动词出现的时候，可以选数词(CD)作为词 b。

以上各选择的动词，在它后面都不能出现“的”，即词性标记为“DEC”或“DEG”的词。

词 a：距离词 b 相对最近的名词（NN、NR、NT）、代词（PN）或量词（M），即在当前子句开头和词 b 位置标记之间寻找词 a。优先选择标记为“NN”和“NR”的名词，其次选择标记为“NT”、“PN”和“M”的词。

词 c：距离词 b 相对最远的名词（NN、NR、NT）、代词（PN）或量词（M），在词 b 位置标记后与子句结束前寻找。选取距离词 b 相对最远的一个即可，没有优先选取原则。

1. 词 a、词 b 和词 c 是否存在，不存在时以“none”作为默认值。逗号 p8 左子句选取的词 a、b、c 为：a_i=银行，b_i=是，c_i=之一。逗号 p8 右子句选取的词 a、b、c 为：a_j=none(默认值)，b_j=是，c_j=银行。左子句和右子句分别以标记 i 和 j 作为区分。
2. 当 a_j 不存在时， 把 a_i 和 b_j 的组合作为一个标签。以逗号 p8 为例，a_i,b_j=银行,是。
3. 当 a_j 不存在时， 把 c_i 和 b_j 的组合作为一个标签。以逗号 p8 为例，c_i,b_j=之一,是。
4. 当 c_i 不存在时， 把 b_i 和 a_j 的组合作为一个标签。逗号 p8 与逗号 p9，不存在该情况。
5. 当 b_i 与 b_j 都存在且二者相同时， 添加 coord=yes 标签。 以逗号 p8 为例，需添加 coord=yes 标签。

3.3.2 当前逗号序号及序号前的逗号分类类别

1. 当前逗号在句子中的序号。从“0”开始计数，如果当前逗号为句子中的第二个逗号，则标记 icom=1。例如逗号[8]的 icom=0。逗号 p9 的 icom=1。
2. 当前逗号序号前的逗号分类类别。在上面的例(4)中，逗号 p8 前面没有逗号。对于逗号 p9，前面有一个逗号 p8，而逗号 p8 根据句法模型的分类结果可知是属于 VP_COORD 分类，则添加 com0=VP_COORD。若例(4)这个句子并没有结束，后面还有逗号，那么下一个逗号要把逗号 p8 与逗号 p9 的分类结果都添加上去，即添加

com0=VP_COORD 和 com1=SB(已知逗号 p9 分类结果为 SB)。

3.3.3 词汇特征

1. 左子句中第一个词及该词词性，用变量 bword_i 与 bpos_i 表示。以逗号 p8 为例，bword_i=中国，bpos_i=NR。
2. 右子句中第一个词及该词词性，用变量 bword_j 与 bpos_j 表示。以逗号 p8 为例，bword_j=也，bpos_j=AD。
3. 取左子句中最后一个词及该词词性，用变量 eword_i 与 epos_i 表示。以逗号 p8 为例，eword_i=之一，epos_i=NN。
4. 左子句中最后一个词(eword_i)是否与左子句中谓语(b_i)相同，若相同，添加 noobj=yes 标签。逗号 p8 和逗号 p9 都不添加该标签。
5. 左子句中第一个词的词性(bpos_i)与最后一个词的词性(epos_i)是否为 P 与 VV，若是，添加 other=yes 标签。逗号 p8 和逗号 p9 都不添加该标签。

表 1 例句(4)中逗号 p8 和逗号 p9 特征选取的值(“-”表示不存在该情况)

特征说明	逗号 p8 的特征	逗号 p9 的特征
左子句词 a	a_i=银行	a_i= none
右子句词 a	a_j=none	a_j=它
左子句词 b	b_i=是	b_i=是
右子句词 b	b_j=是	b_j=拥有
左子句词 c	c_i=之一	c_i=银行
右子句词 c	c_j=银行	c_j=网络
右子句词 a 不存在时，左子句 a_i 与右子句 b_j 的组合作为标签	a_i,b_j=银行,是	-
右子句词 a 不存在时，左子句 c_i 与右子句 b_j 的组合作为标签	c_i,b_j=之一,是	-
左子句词 c 不存在时，左子句 b_i 与右子句 a_j 的组合作为标签	-	-
左右子句词 b 是否相同	coord=yes	-
当前逗号的序号	icom=0	icom=1
当前逗号序号前的逗号分类类别	-	com0=VP_COORD
左子句第一个词及词性	bword_i=中国 bpos_i=NR	bword_i=也 bpos_i=AD
左子句最后一个词及词性	eword_i=之一 eword_i=NN	eword_i=银行 eword_i=NN
右子句第一个词及词性	bword_j=也 bpos_j=AD	bword_j=它 bpos_j=PN
左子句中词 b 与最后一个词是否相同	-	-
左子句首词与尾词是否为 P+VV 结构	-	-

4. 实验结果及分析

4.1 实验数据

实验数据来源于 CTB 6.0，为了便于和 Yang and Xue 的实验结果相比较，本文采用了 Yang and Xue 对 CTB 语料库进行训练数据集和测试数据集的划分方式。详细数据集划分请见表 2。用于训练的逗号有 42497 个，做测试的逗号有 5436 个。

表 2　CTB 6.0 数据集划分

数据	训练	测试
CTB 6.0	81-325,400-454,500-554 590-596,600-885,900 1001-1017,1019,1021-1035 1037-1043,1045-1059,1062-1071 1073-1078,1100-1117,1130-1131 1133-1140,1143-1147,1149-1151 2000-2139,2160-2164,2181-2279 2311-2549,2603-2774,2820-3079	(1-40,901-931 newswire) (1018,1020,1036,1044 1060-1061,1072 1118-1119,1132 1141-1142,1148 magazine) (2156-2180,2295-2310 2570-2602,2800-2819 3110-3145 broadcast news)

4.2 实验结果

Yang and Xue 采用了两种逗号自动分类方法：第一种方法是在加入逗号分类信息的情况下，使用句法分析器对句子重新做句法分析，由此得到逗号分类结果；第二种方法是在句法分析器进行句法分析的基础上，对句法分析过的句子提取特征建立最大熵逗号分类器。其中第二种方法比第一种方法的总体正确率高 1.4%，本文的实验结果将和 Yang and Xue 第二种方法的实验结果进行比较，实验结果及对比在表 3 中展示。可以看到，本文的实验结果在总体正确率上已经非常接近 Yang and Xue 的实验结果，在 ADJ 和 SBJ 两项甚至比 Yang and Xue 的还要高。而 CRF 分类器的效果比最大熵分类器的要好，是由于 CRF 是一种更适合序列标注的模型。由此可见，基于句子的词与词性信息进行逗号自动分类的方法是可行的。

表 3 各分类的实验结果对比

分类	值	Yang and Xue	Maxent	CRF
all	Acc(%)	72.9	66.3	68.1
SB	Prec.(%)	66.2	57.5	60.8
	Rec.(%)	73.1	70.2	72.3
	F.(%)	69.5	63.2	66.1
IP_COORD	Prec.(%)	56.0	55.3	62.1
	Rec.(%)	48.6	27.9	38.5
	F.(%)	52.0	37.1	47.6
VP_COORD	Prec.(%)	68.3	63.8	64.0
	Rec.(%)	78.2	75.6	77.1
	F.(%)	72.9	69.2	70.0
ADJ	Prec.(%)	66.8	61.2	62.0
	Rec.(%)	37.7	50.1	51.7
	F.(%)	48.2	55.1	56.4
Comp	Prec.(%)	91.2	89.6	88.2
	Rec.(%)	92.4	90.8	91.8
	F.(%)	91.8	90.2	90.0
SBJ	Prec.(%)	31.8	53.3	46.2
	Rec.(%)	10	12.5	9.4
	F.(%)	15.6	20.3	15.6
Other	Prec.(%)	85.6	75.8	78.0
	Rec.(%)	84.1	69.6	69.1
	F.(%)	84.8	72.5	73.3

4.2.1 主语连续性

本文对能否自动检测逗号左右两边的主语是否发生变化进行了实验。逗号左右两边子句主语发生变化时，逗号的分类结果属于 SB 或 IP_COORD 分类。主语不发生变化的情况，就是主语保持连续性，且右子句主语因与左子句主语相同而省略，该情况所属分类为 VP_COORD。我们将原先的七类划分为三类，SB 与 IP_COORD 分类归为一类，VP_COORD 分类独自归为一类，其他四类归为 Other 分类。实验结果在表 4 中给出，由此表可知当 SB 分类与 IP_COORD 分类归为同一类时，正确率和召回率相比每个单项都有很大的提高。由此可见，SB 分类与 IP_COORD 分类会发生相互混淆，造成的原因会在错误分析中给出。

表 4 主语连续分类结果

	Maxent			CRF		
%	Prec	Rec	F	Prec	Rec	F
VP_COORD	63.8	75.6	69.2	64.0	77.1	70.0
IP_COORD+SB	72.5	74.0	73.3	72.6	74.8	73.7
Other	84.3	75.6	79.7	86.0	75.6	80.4

4.2.2 语料的差异性

本文实验使用的语料包含了三个来自不同领域的语料，分别是新闻语料、广播稿语料和杂志语料。实验过程中先是把三个语料融合为一个大的语料进行实验，实验结果就是表 3 中展示的。为了考察语料差异性对逗号自动分类结果的影响，接下来又在三个不同的语料上分别做实验，得到的试验结果如表 5 所示。由表 5 可以看出，不同语料的实验结果是存在差异的，在新闻语料中实验效果是最好的，其次是广播稿，最差的是杂志。由于新闻语料是最规范的，所以结果最理想。而杂志语料语言的使用就比较自由、宽松，逗号的使用也就更加随意，得到的结果就最不理想。

表 5 不同语料实验结果

语料	新闻	广播稿	杂志
Maxent(acc%)	70.4	67.7	62.2
CRF(acc%)	74.0	69.7	64.1

4.2.3 逗号作为句子边界

前面相关工作中已经提到了 Xue and Yang(2011)[5]与 Yang and Xue (2012)[8]研究了能否自动检测逗号用于标示句子边界，本文将同样考察该情况。逗号用于标示句子边界的情况就是逗号分类中属于 SB 的分类，所以，只需要把 SB 作为一类，即 EOS；其余的作为一类，即 NEOS。为了与 Yang and Xue 的实验结果相比较，本文也是在新闻语料上做的该实验。实验结果在表 6 中展示。由实验结果可知，CRF 分类器的实验结果非常接近 Yang and Xue 的实验结果，说明本文所提出的方法在自动识别逗号作为句子边界时的效果相对较好。

表 6 逗号作为句子边界

	Yang and Xue			Maxent			CRF		
%	P	R	F	P	R	F	P	R	F
Overall			88.7			83.6			87.5
EOS	63.0	77.9	69.7	50.5	64.9	56.8	60.8	69.4	64.8
NEOS	95.1	91.7	93.4	92.6	87.3	89.9	93.7	91.1	92.4

4.3 错误分析

本文实验是以分词和词性标注的句子作为输入文本来提取特征，通过对错误分类的统计，我们发现了一些存在的问题。存在的问题有：一是相互混淆，比如 SB 与 IP_COORD 的混淆度就比较大；二是偏向性错分，比如一些 OTHER 分类被错分为 ADJ 分类，但很少出现 ADJ 分

类被分为 OTHER 分类。下面将给出详细的分析。

4.3.1 关于 SB 与 IP_COORD

在前面考察主语连续性时，已经提到 SB 与 IP_COORD 分类会发生相互混淆，导致正确率偏低。我们在抽取的一百个逗号分类错误中，有 20 个本该属于 IP_COORD 分类的逗号被错分为 SB 分类，5 个 SB 分类被错分为 IP_COORD 分类。导致这种现状的原因，是因为我们的实验是基于分词与词性标注的句子，没有句法分析，无法得知当前逗号的父节点是否为根节点，以致 SB 与 IP_COORD 分类相互混淆，尤其是 IP_COORD 分类正确率较低。

另外，在这抽取的一百个逗号分类错误中，有 18 个属于 SB 或 IP_COORD 分类的逗号，被错分为 VP_COORD 分类；而这 18 个错分中有 17 个属于子句主语缺省的情况。这种情况属于逗号的右子句主语发生改变，但却省略了主语，在提取特征时右子句词 a 只能保持默认值(none)，导致属于 SB 或 IP_COORD 的分类被错分为 VP_COORD 分类。

4.3.2 关于 COMP 与 ADJ

COMP 分类是指逗号位于句子的谓语和宾语之间。通过下面的三个例子，来阐述说明 COMP 分类。三个例句对应的句法树结构，已在下面的图 3 和图 4 中给出。由逗号分类标准对应的句法树结构可知，逗号 p10 和逗号 p11 属于 OTHER 分类，逗号 p12 属于 COMP 分类。如果把例句(7)的句首加一个“据”字，那么例句(7)将会和例句(6)对应一样的句法树结构，逗号 p12 也将属于 OTHER 分类。事实上逗号 p11 和逗号 p12 的右子句都是“预测”的内容，是谓语的宾语。

(5) 据_P 预测_NN ，_PU p10 今年_NT 全球_NN 经济_NN 增长_NN 幅度_NN 可_VV 达到_VV 百分之四点一_CD 。_PU

(6) 据_P 国家_NN 统计局_NN 预测_VV ，_PU p11 全球_NN 经济_NN 发展_NN 将_AD 给_P 中国_NR 带来_VV 很多_CD 机遇_NN 。_PU

(7) 国家_NN 统计局_NN 预测_VV ，_PU p12 一九九六年_NT 全球_NN 经济_NN 将_AD 继续_VV 保持_VV 增长_NN 。_PU

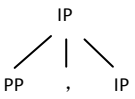

图 3 例句(5)和例句(6)对应的句法树

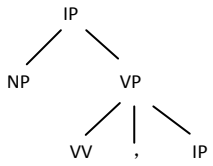

图 4 例句(7)对应的句法树

使用基于句子的分词与词性标注信息做逗号自动分类的方法，就会把一些类似逗号 p11 的情况错分为 COMP 分类。为了避免这种错分，本实验在左子句首词词性为 P 且尾词词性为 VV 时，添加一个 other=yes 标签。这种错分虽然存在，但不是很多。属于 OTHER 分类被误分为 ADJ 分类则比较常见。在抽取的一百个错误的逗号分类结果中，有 24 个 OTHER 类型的逗号被错分为 ADJ 分类，其中有 8 个属于逗号左子句首词词性 P+尾词词性 VV 的情况，也就是类似于逗号 p11 的情况，有 14 个类似于逗号 p10 的情况。由此也可以看出，介词对 ADJ 分类有很大的影响，在这一方面还需要做出更多努力。

5. 结语

本文提出了一种基于分词及词性标注信息的逗号自动分类方法，选择了三组特征，采用了两种有监督的机器学习分类器(即使用最大熵与 CRF 分类器)。实验结果表明，基于词与词性信息的方法在总体正确率上非常接近 Yang and Xue 基于句法信息的方法，可以有效的完成逗号自动分类，证明这种方法是可行的。

尽管实验证明了本文所提出方法的可行性，但是同样存在问题，比如某些逗号分类类别的相互混淆。对于这些不足，我们将会进一步思考更好的解决方法。在下一步的工作中，我

们将会继续立足于逗号，从逗号的角度开展篇章分析的研究，进行深入与细化。

参考文献

[1] Adam L.Berger, Stephen A.Della Pietra, Vincent J.Della Pietra. A Maximum Entropy Approach to Natural Language Processing. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics(ACL). 1996. 39-71

[2] Hen-Hsen Huang and Hsin-His Chen. Chinese Discourse Relation Recognition. [C]//Proceedings of the 5th International Joint Conference on Natural Language Processing 2011, pages1442-1446. 2011. 1442-1446

[3] Meixun Jin, Mi-Young Kim, Dong-Il Kim, and Jong-Hyeok Lee. Segmentation of Chinese Long Sentences Using Commas. [C]//In Proceedings of the SIGHANN Workshop on Chinese Language Processing. 2004.

[4] Nianwen Xue, Fei Xia, Fu-Dong Chiou and Martha Palmer. The Penn Chinese Treeabnk: Phrase Structure Annotation of a Large Corpus. [C]/Proceedings of Natural Language Engineering. 2005. 11(2):207-238.

[5] Nianwen Xue, Yaqin Yang. Chinese sentence segmentation as comma classification. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics(ACL-11). 2011. 631-635

[6] Vanessa Wei Feng, Graeme Hirst. Text-level Discourse with Rich Linguistic Feature. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics(ACL-12). 2012. 60-68

[7] Xing Li, Chengqing Zong, and Rile Hu. A Hierarchical Parsing Approach with Punctuation Processing for Long Sentence Sentences. [C]//Proceedings of the Second International Joint Conference on Natural Language Processing: Companion Volume including Posters/Demos and Tutorial Abstracts. 2005.

[8] Yaqin Yang, Nianwen Xue. Chinese Comma Disambiguation for Discourse Analysis. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics(ACL-12). 2012. 786-794

[9] Yuping Zhou, Nianwen Xue. PDTB-style Discourse Annotation of Chinese Text. [C]//Proceedings of Annual Meeting on Association for Computational Linguistics(ACL-12). 2012. 69-77

[10] Yuqing Guo, Haifeng Wang, and Josef Van Genabith. A Linguistically Inspired Statistical Model for Chinese Punctuation Generation. [C]//Proceedings of ACM Transactions on Asian Language Processing. 2010. 9(2).