

基于逗号的汉语子句识别研究

李艳翠^{1,2} 冯文贺³ 周国栋^{1,†} 朱坤华²

1. 苏州大学计算机科学与技术学院, 苏州 215006; 2. 河南科技学院信息工程学院, 新乡 453003;
3. 河南科技学院人文学院, 新乡 453003; † 通信作者, E-mail: gdzhou@suda.edu.cn

摘要 根据篇章分析的任务和实践, 结合传统研究, 提出汉语的基本篇章单位为子句, 从结构、功能、形式等方面给出其定义。首先分析了逗号与子句的关系, 并在标注语料上进行了基于逗号的汉语子句识别研究。在标注了 CTB6.0 中 1348 个句子的 2171 个逗号可否作为子句边界的语料上, 抽取句法、词汇、长度等特征进行实验, 子句识别准确率为 90%。然后利用信息增益选出贡献最大的 8 个特征, 使用它们也可获得较高的子句识别准确率。最后仅使用词法信息, 子句识别准确率可达 84.5%。实验证明子句的定义合理, 基于逗号的子句识别在理论上和实验上均可行。

关键词 逗号; 汉语子句; 子句识别

中图分类号 TP391

Research of Chinese Clause Identificiton Based on Comma

LI Yancui^{1,2}, FENG Wenhe³, ZHOU Guodong^{1,†}, ZHU Kunhua²

1. Department of Computer Science and Technology, Soochow University, Suzhou 215006; 2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003; 3. School of humanities, Henan Institute of Science and Technology, Xinxiang 453003; † Corresponding author, E-mail: gdzhou@suda.edu.cn

Abstract According to the task of Chinese discourse analysis and practice, combined with traditional study, the authors propose clause as basic discourse unit and give its definition from the structure, function, form etc. The authors analyse the relationship between the comma and clause, and research clause identification using comma on annotation corpus. The corpus labele whether each comma can be regarded as clause boundary information extract from CTB6.0, and have total of 2171commas in 1348 sentences. The authors extract syntax, vocabulary, length features for experiment, and clauses identification accuracy can reach 90%. Eight greatest contribution features are chosen by information gain, they can obtain high clauses identification accuracy. Finally only using morphology feature accuracy can reach 84.5%. Experiments show that the definition of clause is reasonable and identification clause based on the comma is feasible.

Key words comma; Chinese clause; clause identification

自然语言的单位由小到大可以分为字、词、短语、句子和句群, 最后形成篇章(discourse)。篇章有时也称语篇或话语, 通常指由一系列连续的子句、句子或句群构成的语言整体单位。篇章内部具有连贯性, 其内部的各级单位一般是描述同一个主题或同一种情境的。这种连贯性和主题同一性通常由篇章内部子句、句子或句群间的层次结构和语义关系

体现出来, 只有揭示出其中的层次结构及语义关系, 才能对篇章有一个总体把握。

篇章分析是自然语言处理的一个核心问题, 也是近几年的一个研究热点和难点。篇章分析有助于提高自动文摘、问答系统和机器翻译等应用系统的性能。篇章分析的各项工作中, 基本篇章单位识别是一项基础工作, 只有识别了基本篇章单位, 才能

863 计划(2012AA011102)和国家自然科学基金(90920004)资助

收稿日期: 2012-05-30; 修回日期: 2012-08-15; 网络出版时间: 2012-10-26 17:04

网络出版地址: <http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.007.html>

在此基础上做进一步的结构分析。

篇章分析中一般以小句(clause, 也称子句)为基本篇章单位。不过对于什么是小句, 目前并没有完全一致的看法^[1]。生成语法理论认为小句指大于短语的句法单位, 包括传统语法的从句和句子, 还包括一些小句(small clause)。系统功能语法的小句则包括限定小句和非限定小句。除以上的小句所指外, “类型学”中, 有人还把和关系小句具有同样修饰功能的限制性修饰形容词也归为一种关系小句。汉语研究中, 最早提出“小句”说法的是刘复(《文法通论》, 1920), 最早把“小句”当成一级句法单位的是吕叔湘^[2]。吕叔湘的看法是: 小句一般是一个主谓短语; 也常常是一个动词短语(包括只有一个动词的情况); 少数情况下是一个名词短语(包括只有一个名词短语的情况)。赵元任^[3]认为充当句子成分的主谓短语、并列复句中的分句(复合句中的小句)和偏正复句中的分句(从属小句)以及单句都是小句。屈承熹^[4]认为主谓结构都可以看做小句。邢福义^[5]认为“小句是最小的具有表述性的语法单位”, 外延上包括单句, 也包括结构上相当于或大体相当于单句的分句, 但不包括充当句子成分的主谓短语。储泽祥^[6]认为“短语+语调+语用效应”就是小句。可知如何定义小句是篇章分析的一个基本问题。

书面篇章中, 标点符号是篇章单位分割的重要标志。在自然语言处理的不同任务中都能看到引入标点符号的研究。在机器翻译方面, 黄河燕等^[7]利用标点符号和邻近的关系代词配合, 从而把复杂句子切分成多个独立的简单句, 进而提高机器翻译的性能; 在句法分析方面, 李幸等^[8]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^[9]给出利用逗号分割汉语长句的方法, 利用谓词等特征进行分割, 准确率为 87.1%, 长句分割后依存句法分析准确率提高 9.6%。Xue 等^[10]给出一种利用逗号进行汉语句子分割的方法, 自动识别汉语句子中表示句号功能的逗号, 识别的准确率接近 90%。可知基于标点符号进行篇章单位切分是一个常见的方法。汉语篇章中分号、句号、叹号等一般可直接作为篇章单位的切割标志。问题最大的是逗号与汉语篇章单位的关系, 逗号是汉语子句切分的重要标志之一, 但也有很多逗号切分的语言单位并非子句。综上所述, 可知如何处理逗号与

子句的关系是汉语子句识别的一个关键问题。

本文结合已有研究和篇章分析实践, 定义了汉语的基本篇章单位——子句, 探讨它与重要标点(逗号)的关系, 基于逗号进行汉语子句识别研究。

1 逗号和子句

1.1 子句定义

自然语言处理中, 一般把理论语言学中的“小句”称为子句, 本文采用一般的称法。如前所述, 目前对子句并没有统一定义。本文根据篇章分析任务和语料库建设实践, 结合传统研究, 对子句做如下定义。

子句是篇章分析的基本单位, 含传统单句和复句中的分句。结构上, 子句至少包含一个谓语部分, 至少表达一个命题; 功能上, 子句对外不作为其他子句结构的语法成分, 子句和子句间发生命题关系; 形式上, 子句间一定有标点分割, 通常是逗号、分号和句号等。实际语料中, 一些与典型子句在结构、功能、形式上类似的传统所谓短语在特定条件下也作为子句处理。

根据以上定义, 例 1 给出一个段落的子句切分。为便于描述, 各子句开头用大写英文字母标记。

例1 A 浦东开发开放是一项振兴上海, 建设现代化经济、贸易、金融中心的跨世纪工程, B 因此大量出现的是以前不曾遇到过的新情况、新问题。C 对此, 浦东不是简单的采取“干一段时间, 等积累了经验以后再制定法规条例”的做法, D 而是借鉴发达国家和深圳等特区的经验教训, E 聘请国内外有关专家学者, F 积极、及时地制定和推出法规性文件, G 使这些经济活动一出现就被纳入法制轨道。(chttb_0001)

例 1 中, 各子句均包含至少一个谓语部分, 表达至少一个命题, 对外发生命题关系, 形式上用逗号或句号分开。值得指出的是, 子句 C 中包含的语言片段“干一段时间, 等积累了经验以后再制定法规条例”虽然内部结构具备一个子句的形式, 但对外功能上作为一个子句的成分, 不与其他子句发生命题关系。该片段由一个引号标记, 外部不由逗号等切割也显示了其非子句的地位。全面考虑, 可以认定该语言片段不是一个篇章单位, 而只是一个子句内部的成分。

可知判定一个语言片段是否为子句, 需要综合该片段语义和语法两方面的内部结构、外部功能及

一些形式标志等因素全面考虑。子句判定是标注语料库建设的关键问题,为获得一个标注一致的语料库,我们制定了详细的标注规范,用以指导语料库标注工作。考虑到本文主题,这里不做过多陈述。

1.2 逗号和子句识别的关系

由子句定义可知,标点符号是子句的重要形式标志。例 1 显示,根据某些标点(如句号)可直接判定其所切割的语言片段为子句,而另外一些标点(如顿号)所切割的语言片段则绝对不可能是子句,还有的标点(如逗号)切割的语言片段有的是子句,有的不是子句。利用标点进行子句切分是一个常见的方法^[8-9]。不同语言书面语的标点符号用法规定并不相同,不同应用进行句子切分的粒度也不一样,因此利用标点进行子句识别所面临的问题也不完全相同。

文献[11]指出,汉语的常用标点符号有 16 种,分点号和标号两大类。点号的作用在于点断,主要表示说话时的停顿和语气。点号又分为句末点号和句内点号。句末点号用在句末,有句号、问号和叹号 3 种,表示句末的停顿,同时表示句子的语气。句内点号用在句内,有逗号、顿号、分号和冒号 4 种,表示句内的各种不同性质的停顿。标号的作用在于标明,主要标明语句的性质和作用。常用的标号有 9 种:引号、括号、破折号、省略号、着重号、连接号、间隔号、书名号和专名号。可见,汉语书面语中和子句边界有关系的是点号,其中句号、问号、叹号和分号一定表示子句边界;顿号表示句子内部并列词语之间的停顿,所以一定不是子句边界;逗号和冒号可能是,也可能不是子句边界。

不同标点使用频率并不一样。表 1 给出汉语中可能作为子句边界标点的句号、问号、叹号、分号、逗号和冒号等的使用频率,数据统计基于汉语宾州树库(CTB6.0)。

表 1 CTB6.0 中子句边界标点使用频率比较
Table 1 Comparasion of clause boundary punctuation
frequence in CTB6.0

标点	使用频率/%
句号	27.36
问号	1.37
感叹号	0.56
冒号	1.69
分号	1.85
逗号	67.17

表 1 显示,一定是子句边界的句末点号(句号、问号、叹号)和句内点号(分号)共占 31.14%,而可能是子句的逗号却占高达 67.17%的比例。据此可知,子句识别的关键问题是判断逗号是否为子句边界。

需要结合逗号的具体用法来对其作为子句边界的情况进行处理。文献[11]列举了逗号的如下一些基本用法。

1) 句子内部主语与谓语之间如需停顿,用逗号。例如:我们看得见的星星,绝大多数是恒星。

2) 句子内部动词与宾语之间如需停顿,用逗号。例如:应该看到,科学需要一个人贡献出毕生的精力。

3) 句子内部状语后边如需停顿,用逗号。例如:对于这个城市,他并不陌生。

4) 复句内各分句之间的停顿,除了有时要用分号外,都要用逗号。例如:据说苏州园林有一百多处,我到过的不过十多处。

根据本文对子句的定义可知,前 3 种情况下逗号都是子句内部标号,不是子句边界标号,第 4 种情况下逗号是子句边界标号。

综合以上分析可知,逗号对子句分割意义重大。识别逗号是否为子句边界,其实就是识别逗号是否表示复句内各分句之间的停顿。下面针对逗号和子句识别问题进行实验。

2 实验语料

CTB6.0 中共有 2036 个文档,其中前 890 篇文档标注有段落和句子信息,共 18782 个句子,35510 个逗号,4330 段,平均每篇文档 4.8 段,每段 4.3 个句子,每个句子约 2 个逗号。

本文标注子句边界信息所用语料是 CTB6.0 中的前 100 篇文档(chtb_0001~chtb_0100),共 1348 句。对语料中的每个逗号均标注其是否为子句边界信息,“TYPE=Y”表示逗号为子句边界,“TYPE=N”表示逗号非子句边界,标注实例如下。

浦东开发开放是一项振兴上海〈B TYPE=N〉,〈/B〉建设现代化经济、贸易、金融中心的跨世纪工程〈B TYPE=Y〉,〈/B〉因此大量出现的是以前不曾遇到过的新情况、新问题。对此〈B TYPE=N〉,〈/B〉浦东不是简单的采取“干一段时间〈B TYPE=N〉,〈/B〉等积累了经验以后再制定法规条例”的做法〈B TYPE=Y〉,〈/B〉而是借鉴发达国家和深圳等特区的经验教训〈B TYPE=Y〉,〈/B〉聘请国内外有关专家学者〈B TYPE=Y〉,〈/B〉积极、及时地制定和推出法规性文件〈B TYPE=Y〉,〈/B〉使这些经济活动一出现就被纳入法制轨道。

CTB6.0 前 100 篇文档共有 2171 个逗号, 其中标注为正例的逗号有 1355 个, 标注为负例的逗号 816 个。采用手工方法标注逗号是否为子句边界信息, 由一个人独立完成标注。标注时严格按照 1.1 节子句定义进行, 语料的一致性较好。标注完成后, 采用基于特征的机器学习方法, 验证本子句定义方法和基于逗号的子句识别方法的结果如何。

3 特征选择

自动识别逗号是否为子句边界的研究较少, Jin 等^[9]给出一种基于逗号的汉语长句分割方法, 分割的是 IP 和 IP 之间以及短语和 IP 之间的标点, 短语和短语之间的标点不做分割(包括 VP 和 VP 之间的标点)。他们利用标准句法树, 根据分割原理抽取训练和测试样例, 采用谓词及其他相关特征判断逗号可否进行句子分割。Xue 等^[10]给出一个逗号是否为句子边界的自动识别方法, 所用语料为 CTB6.0。他们采用一个启发式规则, 自动将语料中表示句号功能的逗号标注为正例, 其他逗号标注为负例, 然后抽取一些特征, 对每个逗号生成样例进行实验, 结果表明基于逗号的句子识别准确率接近 90%。文献[9-10]识别的是句子, 在句法树中有明显的 IP 标示。本文所识别的是子句(根据定义可知至少包含

IP 和 VP, 主要从语义上进行定义), 粒度要细的多。本文参考文献[9-10]的特征, 加上针对本子句识别任务的一些简单特征, 抽取每个标点的特征进行实验, 所用特征及说明见表 2。

例 1 中子句 A 共有两个逗号, 根据表 2 所述特征, 抽取第 1 个逗号和第 2 个逗号相应特征值实例, 一行表示一个逗号的特征实例。实例中第 1 行表示逗号在整个语料中的顺序号, 第 2 行表示逗号是否为子句边界, +1 表示标点为子句边界, -1 表示非子句边界。抽取逗号特征实例时用到子句 A 的句法信息, 相应的句法树如图 1 所示。

例 1 中第 1 个和第 2 个逗号最终得到的特征实例如下。

1 -1 F1_P_1=NR F1_W_1=浦东 F1_P_2=NN F1_W_2=开发 F1_P_3=NN F1_W_3=开放 F2_P_1=NR F2_W_1=上海 F2_P_2=VV F2_W_2=振兴 F2_P_3=M F2_W_3=项 F3=VV F4=建设 F6=VP F7=VP F8=VP+VP F9=VP+VP+VP F10_1=yesVC F10_4=yesVV F11_4=yesVV F15=, +, +。F17>7 F18=8 F19=yesNP F22=NR+NR F23=浦东+上海

2 +1 F1_P_1=VV F1_W_1=建设 F1_P_2=NN F1_W_2=现代化 F1_P_3=NN F1_W_3=经济 F2_P_1=NN F2_W_1=工程 F2_P_2=JJ F2_W_2=跨世纪 F2_P_3=DEC F2_W_3=的 F3=VV F3=AD F4=因此 F5_2=因此 F6=IP F7=IP F8=IP+IP F9=IP+S+IP F10_1=yesVV f11_1=yesVV f11_2=yesVC F15=, +, +。F17=1 F22=VV+NN F23=建设+工程 F24=DEC

表 2 所用特征及说明表

Table 2 Feature remark table

编号	特征	说明
1	F1_P_N F1_W_N F2_P_N F2_W_N	从逗号到前一逗号或句首的范围内前面 N 个词的词性及词, 后面 N 个词的词性及词
2	F3 F4	逗号之后第一个词的词性和词
3	F5_1 F5_2	逗号前后包含的连接词表中的词(连接词表参考文献[12]中的连接词构建)
4	F6 F7 F8	逗号左右兄弟及其组合的句法信息标记
5	F9	逗号左右兄弟及逗号父亲句法信息组合
6	F10_1 F10_2 F10_3 F10_4 F10_5 F11_1 F11_2 F11_3 F11_4 F11_5	从逗号到前一逗号或句首是否包含是(VC)、表语形容词(VA)、有(VE)、其他动词(VV)或从属连词(CS); 从逗号到后一逗号或句尾是否包含是(VC)、表语形容词(VA)、有(VE)、其他动词(VV)或从属连词(CS)
7	F12 F13 F14	逗号的父节点是否为并列的 IP 结构、逗号是否为第一层子节点、逗号父节点是否是第一层并列 IP 结构
8	F15	逗号所在句子标点的集合
9	F16 F17	逗号到前一标点或句首句子长度是否小于 5, 逗号左右两边句子长度差是否大于 7
10	F18	逗号所在句法树层次
11	F19 F20 F21	逗号的父节、左兄弟节点、右兄弟节点是否为 NP
12	F22 F23	从本标点到一个标点或句首的句子中, 第一个词及最后一个词的词性及词的组合
13	F24	逗号前后所包含相同词的词性信息

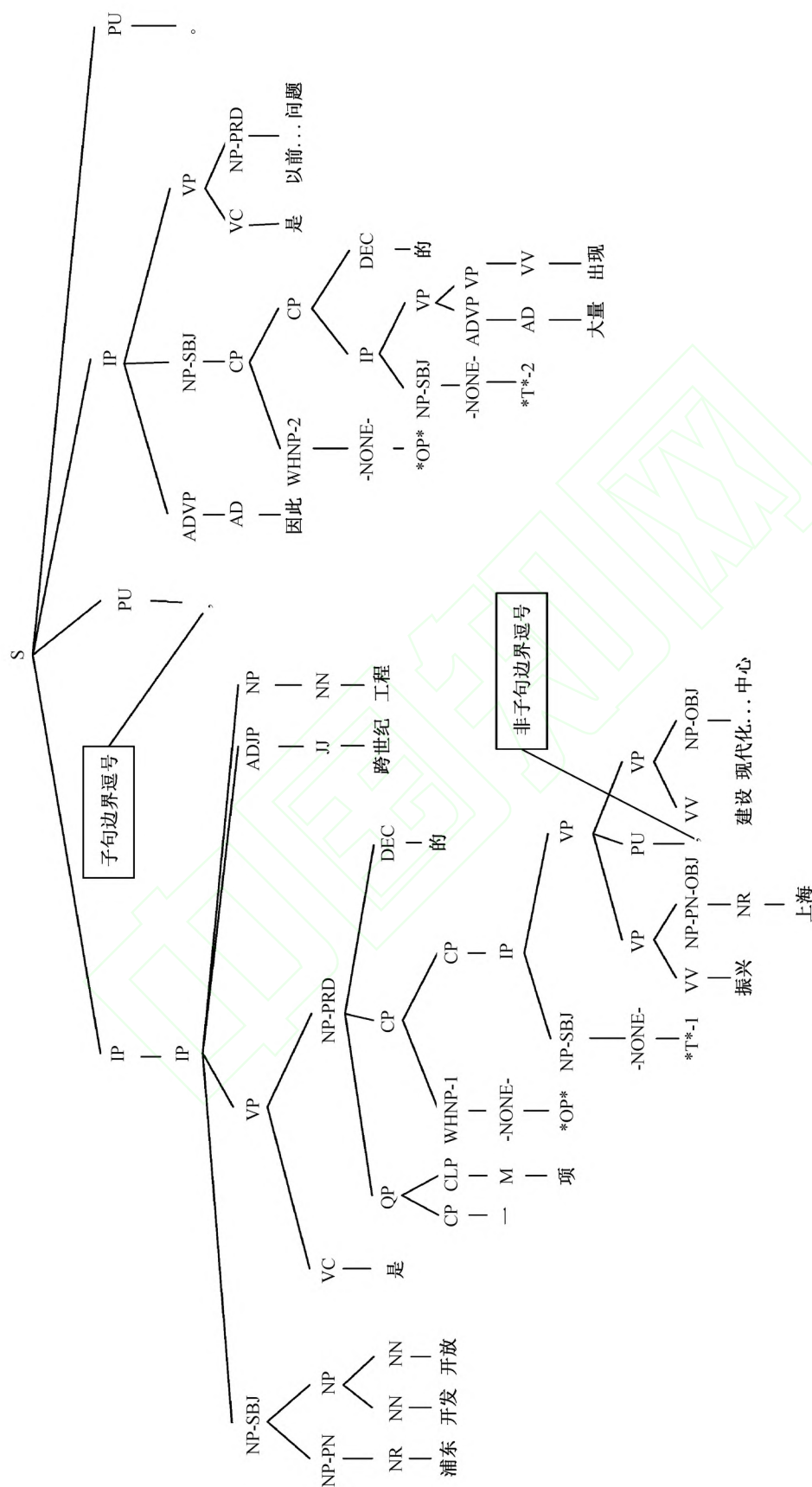


图 1 例 1 中第 1 个句子语法树信息

4 实验结果及分析

4.1 所有特征测试结果

采用第3节提出的全部特征生成逗号对应的特征实例集,然后使用Mallet工具包^[13]中的3个分类器——决策树、最大熵和贝叶斯分别进行实验,使用10次10折交叉验证,实验结果如表3所示。

由表3可知,上述3个分类器中最大熵的效果最好。采用CTB6.0提供的标准句法树,最大准确率为92.8%;采用Berkeley自动句法分析树,准确率是89.9%。对于逗号为子句边界的情况, F 值标准句法树最高94.5%,自动句法树最高91.6%。对于逗号非子句边界的情况,标准句法树 F 值最高90.1%,自动句法树 F 值为86.5%。从表3可以看出,逗号非子句边界的情况比逗号为子句边界的情况识别效果差,归纳主要原因如下:训练样例不均衡,正负比大约为1:1.7;逗号是非子句边界的情况比较复杂,如有子句内部主语与谓语之间的情况,子句内部动词与宾语之间的情况,判断起来比较困难,

且较难找到非常有效的特征。

4.2 单组特征贡献测试

第3节给出了13组特征,这些特征到底有什么作用?下面通过实验,分析不同特征组对系统的贡献。

由于4.1节实验中没有指定训练集和测试集,所以数据都是随机分割的。不同的分割方法可能会对实验结果产生影响,因此该方法很难得出特征的实际贡献。本文在验证特征贡献时,将100篇文档产生的2171个样例分为训练集和测试集两部分:训练集使用前90篇文档,共1960个样例,其中正例1214个,负例746个;测试集使用后10篇文档,共211个样例,其中正例141个,负例70个。分类器使用中表现效果最好的最大熵,实验结果如表4所示。

从表4可以得知,在所有特征中减去某组特征,分别采用标注句法树和自动句法树,结果变化均不大;分别减去(F1 F2)、(F6 F7 F8)、(F10 F11)和(F12

表3 使用全部特征子句边界识别结果

Table 3 Clauses boundary recognition results using all features

分类器	标准句法树			自动句法树		
	准确率/%	$F_1(+)$	$F_1(-)$	准确率	$F_1(+)$	$F_1(-)$
最大熵	92.8	94.5	90.1	89.9	91.6	86.5
决策树	91.7	93.7	89.3	89.4	91.6	86.5
贝叶斯	90.6	92.4	87.1	87.5	90.2	83.0

表4 不同特征组的贡献度

Table 4 Contributions of different feature group

	标准句法树			自动句法树		
	准确率/%	$F_1(+)$	$F_1(-)$	准确率/%	$F_1(+)$	$F_1(-)$
all	91.0	92.7	88.2	89.6	91.5	86.1
-(F1 F2)	89.1	91.1	86.1	89.1	91.2	85.7
-(F3 F4)	90.8	92.5	88.0	89.6	91.7	86.1
-(F5)	91.0	92.7	88.2	89.6	91.7	86.1
-(F6 F7 F8)	90.2	92.5	87.2	88.5	90.9	84.6
-(F9)	90.6	92.2	88.2	89.6	91.7	86.1
-(F10_1~5 F11_1~5)	90.0	92.2	88.4	88.7	90.8	86.0
-(F12 F13 F14)	90.8	92.5	88.0	88.7	90.8	85.2
-(F15)	91.0	92.7	88.2	89.3	91.4	85.8
-(F16 F17)	91.0	92.7	88.2	89.0	91.1	85.4
-(F18)	90.8	92.5	88.0	89.3	91.4	85.8
-(F19 20 21)	90.5	92.3	87.5	89.1	91.2	85.7
-(F22 F23)	90.7	92.5	87.7	89.3	91.4	85.8
-(F24)	91.0	92.7	88.2	89.6	91.5	86.1

F13 F14), 实验结果略有下降。实验结果说明这 4 组特征有一定的作用, 其他的几组特征作用并不明显。为找出所有特征中较有用的特征, 下面采用信息增益的方法进行特征选择。

4.3 使用信息增益进行特征选择

由 4.2 节的实验可知, 使用单个特征的实验方法很难直接判断哪个特征有用, 因此本文参考文本分类中特征抽取的方法, 采用信息增益(information gain, IG)计算有效的特征。对文本分类任务而言, 计算信息增益是针对每个特征项而言的, 对于特征项 t 和文档类别 c , 通过计算 c 中特征项 t 出现与否的文档数来衡量特征项 t 对类别 c 的信息增益, 当 IG 小于预定值时, 特征 t 从特征集中移去, 进而获得文本分类的有效特征。IG 的定义^[14]如下:

$$G(t) = -\sum_{i=1}^m p(c_i) \log P(c_i) + \\ p(t) \sum_{i=1}^m p(c_i | t) \log p(c_i | t) + \\ p(\bar{t}) \sum_{i=1}^m p(c_i | \bar{t}) \log p(c_i | \bar{t}), \quad (1)$$

$p(c_i | t)$ 表示文本中出现特征 t 时文本属于 c_i 的概率, $p(c_i | \bar{t})$ 表示文本中不出现特征 t 时文本属于 c_i 的概率, $p(c_i)$ 表示类别出现的概率, $p(t)$ 表示 t 在整个文本训练集中出现的概率。

参考文本分类任务, 针对本文中的子句识别任务, c 为正负两类, t 为上面所提到的特征取值, 计算概率时将一个实例当做一个文本来计算, 文本的类别即为实例的类别。显然, 某个特征项的信息增益值越大, 表示其贡献越大, 对分类也越重要。本实验选择信息增益大于 0.5 的特征值。

根据式(1)计算特征值的信息增益, 对于正例而言, 重要的特征是 F2 F3 F4 F6 F7 F8 F9 F23 F24, 对于负例重要的特征是 F2 F3 F4 F6 F7 F8 F9 F23, 综合后取特征 F2 F3 F4 F6 F7 F8 F9 F23 F24 作为选择后的特征, 在标准句法树和自动句法树上进行实验, 使用 10 次 10 折交叉验证, 结果如表 5 所示。

比较表 5 和表 3 可知, 使用全部 24 个特征和通

过信息增益抽取出来的 9 个特征结果相当。可见对子句边界识别任务而言, 子句开始和结尾的词及其组合, 子句后逗号左右兄弟及父节点的句法信息对子句识别非常重要。由于本实验结果依赖句法分析结果, 很多时候在句法分析前, 就必须考虑句子的正确切分。因此本文将 24 个特征中的有关句法特征去掉, 讨论基本特征对子句识别的贡献。

4.4 使用非句法特征进行子句切分

从前面的实验可知, 句法信息对子句的正确切分帮助很大, 虽然我们可以使用自动句法分析结果进行子句切分, 但如果能避开句法, 在词法层面上实现子句的切分, 对自然语言处理的意义将是很大的。本文将表 1 中和句法相关的特征去掉, 利用 F1 F2 F3 F4 F5 F10 F11 F15 F16 F17 F22 F23 F24 进行实验, 采用 Berkeley 句法分析器产生的分词及词性标注结果, 结果如表 6 所示。

使用非句法信息进行子句边界识别的准确率最高是 84.6%, 比采用句法信息的最好结果低 4.5%, 说明句法信息对子句识别非常重要。但有些任务需要在句法分析前识别子句, 使用简单特征的结果可以为其他自然语言任务(如句法分析)提供帮助。

实验中有约 10% 的情况子句识别错误, 到底错误原因在哪里? 下面对其进行分析。

4.5 错误分析

4.5.1 负例错误识别为正例

从上面实验可以看出, 负例识别的效果较差, 错误识别的逗号有以下情况: 子句内部主语与谓语之间的逗号; 子句内部动词与宾语之间的逗号; 子句内部状语之间的逗号。识别错误的逗号用下划线标示, 具体情况如下。

例 2 出口快速增长, 成为推动经济增长的重要力量。(chth_0097)

例 3 确立了以资源换技术, 以产权换资金, 以市场换项目, 以存量换增量的利用外资新思路。(chth_0091)

表 5 使用选择后特征的子句边界识别结果
Table 5 Clauses boundary recognition results using selected features

分类器	标准句法树			自动句法树		
	准确率/%	$F_1(+)$	$F_1(-)$	准确率/%	$F_1(+)$	$F_1(-)$
最大熵	92.2	94.0	89.5	89.1	95.6	85.8
决策树	92.0	93.0	89.1	89.1	95.1	86.4
贝叶斯	89.7	91.3	86.1	85.3	88.6	79.4

表 6 使用非句法特征的句子边界识别结果
Table 6 Clauses boundary recognition results not using syntactic features

分类器	准确率/%	$F_1(+)$	$F_1(-)$
最大熵	84.6	88.5	78.6
决策树	82.6	86.9	74.4
贝叶斯	80.5	85.5	72.6

例 4 天津港保税区投入运行五年来,已建成了中国第一货物分拨中心,具备了口岸关的功能,开通了天津港保税区经西安、兰州到新疆阿拉山口口岸的铁路专用线。(chtb_0099)

例 2 中的逗号前面是整个句子的主语,逗号不能作为子句边界,但句法分析将“出口快速增长”分析为 IP,导致此逗号没有正确分类;只使用词法信息,其中包含出口(NN)和增长(VV),也较易识别为正例。例 3 中的逗号是句内动词和宾语之间的逗号,但系统识别时将其识别为子句,因为现在没有体现他们是句子宾语的信息。例 4 中的“天津港保税区投入运行五年来”是整个句子的时间状语,逗号表示句子内部状语之间的停顿,但“天津港保税区投入运行五年”是个子句,从而导致识别错误。

4.5.2 正例错误识别为负例

例 5 内地经济长期稳定地增长,香港经济将从充满活力的内地经济中获益。(chtb_0093)

例 5 中的逗号前的增长是动词,逗号后的香港是名词,导致分类器没有正确分类。

5 结语

本文给出了篇章分析的基本单位——子句的定义。子句识别和逗号关系密切,本文标注了 100 篇 CTB6.0 语料中逗号可否作为子句边界的信息,然后抽取词、句法和位置等信息进行识别,结果表明本文子句自动识别准确率可达 90%。本文初步给出的特征较多,又通过实验找到 8 个比较有效的特征,使用这 8 个特征子句识别的结果和使用全部特征相

差不多,说明本文抽取出的特征较有效。下一步的工作是进一步提高子句识别的性能,以便为篇章分析中结构树的构建提供基础。

参考文献

- [1] 王文格. 现代汉语小句的研究现状及存在的问题. 汉语学习, 2010(1): 67-76
- [2] 吕淑湘. 汉语语法分析问题. 北京: 商务印书馆, 1979
- [3] 赵元任. 汉语口语语法. 北京: 商务印书馆, 1979
- [4] 曲承熹. 汉语篇章语法. 潘文国, 译. 北京: 北京语言大学出版社, 2006
- [5] 邢福义. 小句中枢说. 中国语文, 1995(6): 420-428
- [6] 储泽祥. 小句是汉语语法基本的动态单位. 汉语学报, 2004(2): 48-55
- [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. 中文信息学报, 2002, 16(3): 1-7
- [8] 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. 中文信息学报, 2006, 20(4): 8-15
- [9] Jin M X, Kim M Y, Kim D, et al. Segmentation of Chinese long sentences using commas // Proceedings of 3rd ACL SIGHAN Workshop. Barcelona, 2004: 1-8
- [10] Xue Nianwen, Yang Yaqin. Chinese sentence segmentation as comma classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, 2011: 631-635
- [11] 国家语言文字工作委员会. GB/T 15834-1995, 标点符号用法. 中国标准出版社, 1995
- [12] 邢福义. 汉语复句研究. 北京: 商务印书馆, 2003
- [13] McCallum A K. Mallet: a machine learning for language toolkit [CP/OL]. (2002)[2012-02-28]. <http://mallet.cs.umass.edu>
- [14] Yang Y, Pederson J O. A comparative study on feature selection in text categorization // Proceedings of the 14th International Conference on Machine learning (ICML'97). Nashville: Morgan Kaufmann Publishers, 1997: 412-420