

汉语事件描述小句自动识别^{*}

陈丽欧¹, 周强²

¹清华大学计算机系, 北京 100084

²清华大学信息技术研究院语音和语言技术中心, 北京 100084

¹chouou@foxmail.com, ²zq-lxd@mail.tsinghua.edu.cn

摘要: 本文提出了一种汉语事件描述小句的自动识别方法, 通过对事件描述小句边界分布情况的分析, 将该识别任务转化为对句中特殊符号分类的任务。利用最大熵分类器, 选择两类有效的特征, 重点解决对非结句点号的分类, 并在后处理阶段中总结了对提高识别性能有帮助的规则, 最终在测试集上获得了 79.98 的 F1 值。最后, 总结了识别方法的思想, 分析现有处理系统的不足之处, 并提出了一些展望。

关键词: 点号, 分类, 后处理

Automatic Identification of Chinese Event Descriptive Clause

Chen Liou¹ Zhou Qiang²

¹Department of Computer Science and technology, Tsinghua University, Beijing 100084, China

²Center for Speech and Language Technologies, Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China

¹chouou@foxmail.com, ²zq-lxd@mail.tsinghua.edu.cn

Abstract: We propose an automatic method to identify Chinese Event Description Clause. By analyzing the boundary distribution of clauses, we formulate this identification task as a classification of special symbols. The maximum entropy classifier is trained and two kinds of useful features and their combinations are used to classify the Non-End Symbol. After identifying all clauses, a rule-based post-processing phase for improving the clause recognition performance is included, and ultimately F1=79.98 result is obtained on the test set. Finally, we summarize the method, analysis some deficiencies in current system and give the future research directions.

Key words: punctuation mark, classify, Post-processing

1 概述

汉语事件描述小句 (EDC) 定义为以逗号、分号、句号、问号等点号分隔而形成的词语序列, 它是包含完整事件内容信息的最小单元。在 EDC 的基础上可以做进一步的句法分析和语义理解, 对自然语言处理具有重要的意义, 因此迫切需要对 EDC 识别做深入的研究。同时, 由于汉语 EDC 平均长度较长 (9 个词以上), 内部组成复杂, 且点号的使用非常灵活, 又导致 EDC 的识别具有一定的挑战性。

对 EDC 的自动识别, 国内外的相关研究不多。Steven Abney[1] 提出了一种子句过滤器; Leffa^[2] 实现了一种基于规则的英语及葡萄牙语文本中子句识别方法; Orasan^[3] 在 Susanne 语料库上完成一

^{*} 本研究得到国家自然科学基金项目 (编号: 60573185, 60873173) 和国家高科技研究发展计划 (编号 2007AA01Z173) 资助。

种基于记忆学习方法的子句识别系统,该系统还包括一个基于规则的后处理阶段;CoNLL-2001^[4]也对英语子句识别任务进行过评测。

英语子句识别基本包括三个阶段,子句起点识别、终点识别和完整嵌套结构识别。本文的EDC识别不考虑小句内的嵌套结构,仅从输入的经过分词及词性标注的句子当中识别出所有上层EDC的边界(起点、终点)。

现有的英文子句识别方法通常是基于子句间具有比较明显的先行词这一特征的,而汉语EDC则是以点号作为分隔,子句间没有明显的标记。考虑到汉语EDC的特殊结构,本文将EDC识别任务转化为对句子中可充当EDC边界的符号的识别问题,通过选取分类特征,构造对符号进行分类的分类器,从而识别出相邻两个自由符号之间的EDC,并加入基于规则的后处理步骤,进一步提升识别性能。实验结果表明,这种做法有效可行。

2 设计思路

本文试图将EDC的识别任务视作对句子中**特殊符号**的分类任务,即判断出句子中所有特殊符号是否为**自由符号**。在此,我们定义两个术语:“特殊符号”与“自由符号”。将有可能担当EDC边界的位置或者标点符号称作“特殊符号”;如若这些符号或位置确实为EDC的边界,则称之为“自由符号”。

为判断这种处理思路是否可行,我们对EDC的分布情况进行统计,统计方法如下:

首先,某些句子不需要进行EDC识别,无需启用EDC识别系统,例如:“地理学/n”。这样的句子中词个数不大于2个,且都是特殊词性的词,容易判别。

其次,在需要进行EDC识别的句子中,我们再根据不同EDC边界统计下面几种情况:

1、EDC的边界为句子的起始位置或结束位置,如:[林超/nP 杨吾扬/nP]

2、EDC的边界为点号。为更清晰起见,我们将点号再细分为结句点号(wE)、非结句点号(主要包括冒号(wM)和分隔点号(wP))。因此这类情况又可细分为:

2.1、EDC边界为非结句点号,如:[在/p 地球/n 表面/n 形成/vN 过程/n 中/f],/wP

2.2、EDC边界为结句点号:如:[大陆/nS 的/uJDE 面积/n 几经变迁/v]。/wP

3、EDC的边界为标号,如:[人/n 称/v]:/wM “/wLB [山河/n 依旧/d],/wP [面貌/n 不/dN 改/v]”/wRB。/wE

4、EDC的边界为其他情况,如:[每/rB 年/qT 组织/v 千/m 余/m 名/qN 学员/n][行程/n 800/m 公里/qN][到/v 老区/n 进行/v 野营/v 施教/v]。/wE

统计时使用的语料为CIPS-ParsEval-2009任务^[5]提供的训练数据,数据格式为{[EDC]*},有171个文件(其中23个BAIKE类型、23个HYL类和125个NEWS类别),共14248个汉语句子。本文所用语料,如无特殊说明,均为CIPS-ParsEval-2009任务提供。

统计语料的14248句子中,共198(1.39%)条无需启用EDC识别系统。统计剩余14050个有EDC的句子的66934个EDC边界的分布情况。统计结果如表1所示。

表1 EDC边界情况分布

类型	句子始/末	结句点号	非结句点号	标号	其他
个数	14710	13446	37173	795	810
比例(%)	21.98	20.09	55.54	1.19	1.21

由表 1 看出，21.98%属于边界位于句子始/末的情况，这种情况具有比较明显的特征，可使用简单的规则进行分类；绝大部分（75.63%）属于边界为点号的情况；1.19%属于边界为标号的情况，通常会与点号配合出现，通过点号的识别将能实现对标号的识别。以上三种位置/标点符号即本文所考察的“特殊符号”。通过对这些特殊符号的处理，将能够覆盖 98.79%的边界情况。因此我们将识别的重点放在对各种特殊符号是否可作为 EDC 边界进行二分类的思想是可行的。

剩余的其他没有明显特征的 1.21%边界情况，由于情况十分复杂，且所占比例不大，因此在本文中并没有进行处理。

3 方法实现

3.1 EDC 识别的整体设计

按照将识别问题转化为分类问题的思路，我们设计 EDC 识别的整体过程如图 1 所示。

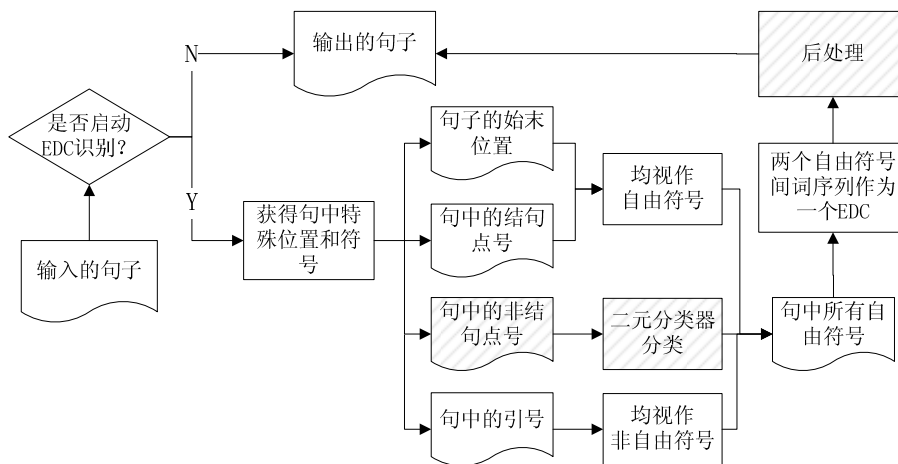


图 1 EDC 识别整体过程

对输入的句子，我们先根据句中词个数及词性两个条件，来决定是否要启动 EDC 识别步骤。

对需要进行 EDC 识别的句子，根据表 1 显示的边界分布情况，我们主要考察这些特殊符号：

1、句子始末

对句子的起始位置，我们认为它可充当第一个 EDC 的左边界；句子的结束位置可充当最后一个 EDC 的右边界，因此，句子的始末位置都被视作自由“符号”。当然，以句子始末作为边界得到的 EDC 有一些特殊情况，这些将在最后的后处理阶段统一处理。

2、结句点号

一般来讲，汉语中的结句点号都作为一个句子的结束标记，因此在本系统中，我们假设所有的结句点号都是自由符号，也就是结句点号都能充当 EDC 的边界。

3、非结句点号

从表 1 可以看出，EDC 的边界中，非结句点号所占的比例最大。且由于其使用很灵活，因此在我们的任务中，重点解决非结句点号的分类问题，我们将使用一个分类器，对每一个非结句

点号进行自由点号/非自由点号的二分类。

4、标号

本文中，我们假设所有的标号都是非自由符号，都不充当 EDC 的边界，仅在后处理阶段中处理一种特殊的可作为自由符号的标号：引号。

以上四类情况，基本上覆盖了 EDC 边界的绝大部分(98.79%)，是本系统重点关注的。

5、其他情况

这种情况虽然所占的比例不大，但情况较复杂，难以找到分界点，因此在本文中不进行处理。

通过对以上各种情况的不同处理，我们将得到句中所有的自由符号，此时，相邻两个自由符号之间的词串将形成一个 EDC，经过后处理步骤的最后调整，可输出 EDC 的分析结果。

3.2 非结句点号的二分类

非结句点号既可以作为 EDC 的边界，还可以用来分隔 EDC 内部各个功能块、分隔各功能块内部的并列成分、分隔复杂从句内部的小句等等，使用非常灵活。CIPS-ParsEval-2009 任务提供的数据中，共 33507 个 EDC，其中有 6558 个 (19.57%) EDC 中都包含非结句点号，因此如果简单的将非结句点号判断为自由或非自由是很不合适的。

为解决此类问题，本文实现了一个分类器，对非结句点号做自由点号/非自由点号进行二分类。这是本文的重要工作及特色所在，将有助于解决对点号是否可作为 EDC 边界的判定问题。

在训练分类器时，最重要的是特征的选择。选择特征时我们基于以下几个方面的考虑：

1、局部语境的词语和词性信息

词语和词性信息是输入句子中可得到的最直观的信息，同时我们认为，点号周围一些特殊的词、词性等会对判断该点号是否自由有一定作用。可使用的词语和词性信息可有：

1.1、待分类点号自身信息

主要是点号的词性。我们认为对分隔点号(wP)和冒号(wM)的处理应该不同，对于冒号(wM)，它在一个 EDC 内部的可能比较大。

1.2、待分类点号附近的词+词性信息

我们认为，词+词性是对一个词的完整描述，因此当选取特征时，也应该将这两部分结合起来作为一个组合特征进行使用。同时，为更全面地利用点号周围的词汇信息，我们设置语境窗口可在[0, 0]-[-4, 4]之间变化，通过实验来选择合适的窗口。

1.3、待分类点号前后相邻点号的信息，主要有：

与相邻点号的距离：我们认为，在逗号/分号充当并列结构的分隔标点的情况下，两个相邻点号之间的间隔词语数目相对来说会较少，因此距离较短。

相邻点号的词以及本点号的词：这个特征可与距离特征配合使用。我们认为，如果相邻点号是逗号/分号，且当前待分类点号也是逗号/分号，且两个点号之间距离较小，则有很大的可能是逗号/分号充当并列结构的情况。

1.4、待分类点号前后相邻标号的信息，主要有：

与相邻标号的距离：这有助于判断由标号对形成的分层次的 EDC。

相邻标号的词性：如果左相邻标号是 wLB，则这个点号在标号对内部，那么很有可能不是自由点号，同理，如果右相邻标号是 wRB，也有很大可能不是自由点号。

以上是词汇方面可以利用的特征。实验结果表明，句子结构方面的特点无法完全通过表面的

词汇体现出来,仅使用词汇特征构造的分类器对并列复句结构及逗号充当并列结构分隔这两种情况的区分性不高,对由点号分隔开的复杂名词短语或句子构成的 EDC 内部主语 (S) 或状语 (D) 的识别性能欠佳。因此仅使用词汇特征是不够的。为此,我们又考虑引入句法方面的信息。

2、局部语境的功能块信息

我们对输入的句子做句法分析,希望能从句法分析结果中提取到有效的、能够解决词汇特征所无法解决的问题的特征。可使用到的功能块信息有:

2.1、点号位置

即待分类点号是否在某个功能块内部。通过句法块分析,能够将某些内部含有点号的复杂结构封装为一个复杂主语或状语块,当点号在块内部时,它作为自由点号的可能性非常小。

2.2、点号周围功能块的信息

词汇特征无法体现整个句子的结构,因而也就不能利用到汉语句子典型句式的特点。为解决这个问题,可以将点号前后的功能块序列信息作为一个特征,我们设置功能块序列窗口在 $[0, 0] \sim [-4, 4]$ 之间变化,通过实验来选择合适的窗口大小。

2.3、点号左相邻区间功能块序列

“左相邻区间”指的是待分类点号与其左相邻点号之间的部分。我们使用这个区间内的功能块序列,期望可以覆盖复杂名词短语/句子做主语/状语的情况。同时,功能块序列还应该包括块之间的功能词(如介词/p、结构助词/uJDE),这样有助于对从句的判断。

2.4、点号左相邻区间内谓语 (P) 块的个数

我们认为,如果点号左相邻区间内没有谓语块,很有可能这个点号前的部分仅是一个完整 EDC 的状语部分,它不是自由点号;如果该区间内有多个谓语块,有可能是点号充当并列结构分隔的情况,也不是自由点号。

以上我们从两个大方面分析提出了一些对非结句点号分类可能有效的特征,将通过实验来选择真正能够提高分类性能的特征集。

3.3 后处理

从图 1 可看出,通过对特殊符号进行分类得到两个自由符号之间 EDC 之后,还会利用一些简单的启发式规则做后处理,将后处理结果作为最终结果输出。

(1) 两个边界内的词是特殊词的后处理:若两个边界内的词是特殊词(连词),则不作为 EDC。

(2) 引号的后处理:如果左引号左边直接相邻冒号(:/wM),且该冒号被分类为自由点号,那么认为相应 EDC 的左边界不是冒号,而是这个左引号;如果右引号右边直接相邻结句点号,那么认为相应 EDC 的右边界不是结句点号,而是这个右引号。

(3) 特殊序号的后处理:对特殊序号(如“①”)统一处理为一个 EDC。

(4) 状语的后处理:如果句中只识别出了两个 EDC,且第一个 EDC 只由一个状语块组成,那么认为这个状语块不该作为一个单独的 EDC,需要将两个合并成一个 EDC。

4 实验结果及分析

4.1 数据划分

4.1.1、针对非结句点号的二元分类器

训练和测试二元分类器时，我们使用 CIPS-ParsEval-2009 评测提供的训练数据，共 171 个文件。按照 80%、10%和 10%的比例，随机将这 171 个文件划分为训练集、测试集和开发集。

4.1.2、整个 EDC 识别系统

对整个 EDC 识别系统做评测时，我们用全部 171 个文件作分类器的训练数据，测试数据采用 CIPS-ParsEval-2009 评测提供的评测数据，共 1 个文件，3751 个句子。

4.2 实验方法

4.2.1、对二元分类器特征选择的实验

训练和测试数据如 4.1.1 中介绍，分类器采用最大熵工具包¹，考虑局部语境的功能块信息时，主要使用了在相同文本的功能块训练库上训练得到的功能块分析器^[6]，其整体识别F1 值约为 85%。使用识别精度（识别正确的点号个数 / 所有待识别的点号个数）评价分类器的性能。

通过该实验，将得到对非结句点号分类有效的特征集。

4.2.2、对整个 EDC 识别系统的评测

使用 4.1.2 中的数据，整个识别过程如图 1 所示，使用 EDC 识别的 F 值来对识别性能进行评价。通过该评测，能更加明确本文提出的方法的识别能力和不足之处。

4.3 实验结果

4.3.1、对二元分类器特征选择的实验

该实验我们采用不断添加各种特征的方法，从中挑出对提高分类性能有帮助的特征，而去掉那些对性能无影响或导致性能降低的特征，最终得到一个特征集合。

加入 3.2 节中介绍的每个特征后分类器识别精度变化如表 2 所示（在开发集上）。

点号周围词信息和功能块信息的语境窗口大小也是通过实验获得的，通过将窗口从[0, 0]向 [-4, 4]之间依次移动，选取使得识别精度最大的窗口值

表 2 特征对识别精度的影响

特征	识别精度 (%)
仅使用点号自身词性	66.19
加入点号周围词+词性（语境窗口大小为[-4, 3]）	70.86
加入点号前后相邻点号信息(点号自身词+相邻点号词+两个点号的距离)	71.80
加入点号前后相邻标号信息(相邻标号的词性+点号与标号间的距离)	72.91
加入点号位置特征	74.42
加入点号周围功能块信息（语境窗口大小为[-4, 4]）	81.45
加入点号左相邻区间的功能块序列	82.25
加入点号左相邻区间内谓语句的个数	82.70

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

4.3.2、整个 EDC 识别系统的评测

对整个 EDC 识别系统评测时，使用两种方式，其一是构造非结句点号的二元分类器时仅加入表 2 中的词汇特征，另一个是在词汇特征的基础上，又加入表 2 列出的句法块特征。

系统的识别性能如表 3 所示。

表 3 EDC 识别系统的性能

系统	F 值 (%)
仅使用词汇特征	72.83
词汇特征+句法块特征	79.98

4.4 实验结果分析

4.4.1、对二元分类器特征选择的实验

随着特征的加入，二元分类器对非结句点号是否为自由点号的分类性能逐渐提高。

在词汇特征中，点号周围的词+词汇特征对识别性能的贡献较大（约 4%），而点号前后其余点号及标号的特征却对性能提高较有限。我们认为，点号周围点/标号特征的引入是为了解决并列结构的识别问题，但简单的距离、词性等信息对并列情况判断的贡献是不够的，更重要的应该是分析句子内部的结构。所以引入句法块特征是很合理的，实验结果也证明了这一点，当加入句法块特征后，识别精度提高了近 10 个百分点。

在句法块特征中，点号周围功能块信息是最具区分性的特征，能提高分类精度近 7%，而左相邻区间内 P 块个数则效果不太显著。我们认为，这是左相邻区间内 P 块个数的特征基本已被点号周围功能块序列这一特征所覆盖的缘故。

对错误文件进行分析后我们发现，加入句法块特征后，有效提高了对 EDC 小句中的主语由复杂名词短语或句子构成，且该主语与其他部分之间由点号隔开的情况的识别性能，并对从句情况的判断有较好的效果。但这些依旧难以解决以下难点：

1、复句层面与单句层面状语的区别问题

本文的 EDC 涉及到两种不同类型的状语，复句层面的状语将作为一个单独的 EDC，而单句层面的状语则和其余部分一起作为一个 EDC。但对于一个状语，很难判断它是仅修饰与它直接相邻的小片段的话，还是修饰了整个一个句子，这涉及到了语义层面的信息。目前的处理中，也没有提取到对这一修饰关系的判断有效的特征。例如：

a、[在/p 地球/n 表面/n 形成/vN 过程/n 中/f]，/wP [大陆/nS 与/cC 海洋/n 的 /uJDE 面积/n 几经变迁/v]，/wP [气候/n 多/m 次/qV 交替/vN]。/wE

b、[在/p 陆地/n 上/f]，/wP 气候/n 从/p 沿海/s 向/p 内陆/nS 呈/v 规律性/n 变化/vN]，/wP [由/p 陆地/n 边缘/n 向/p 内陆/nS 中心/n]，/wP [气候/n 由/p 湿润 /a 、/wD 半/m 湿润/a 、/wD 半/m 干旱/a 到/v 干旱/a]，/wP

a、b 两例中，前例是复句层面的状语，而后例是单句层面。仅靠现在的特征，是难以判断出这两种状语的区别的。

2、由点号分隔的单句并列结构和并列复句的区别问题

如果要判断一个点号是充当单句中并列结构的分隔还是并列复句的分隔，需要综合考虑该点号前后一系列结构的关系，而目前的特征集中，没有对特征进行向后传递，因此对并列情况的处

理也差强人意。例如：

a、[生物/n 由/p 海洋/n 发展/v 到/vB 陆地/n]，/wP [由/p 简单/a 到/v 复杂/a]，/wP [由/p 低级/a 到/v 高级/a]

b、[预计/v 肉类/n 总产/n 增长/v 4%/m ，/wP 奶类/n 增长/v 3%/m ，/wP 禽蛋/n 增长/v 5%/m ，/wP 水产品/n 增长/v 5.3%/m]

以上两例中，a 是并列复句，而 b 是并列结构的单句。要判断 b 中并列结构的情况，需要考虑到“预计”这个动词对后续的“奶类增长”、“禽蛋增长”、“水产增长”等具有控制作用，因此“预计”这个词的特征需要传递到后面的结构中去，而现有的特征则未涉及到这一点。

因此，后续工作我们将着力于解决以上两个难点。

4.4.2、对整个 EDC 识别系统

从表 3 中可以看出，当非结句点号的识别性能提高时，相应整个系统的性能也会提高，整个 EDC 识别系统的性能是与非结句点号的分类性能呈正相关的。因此，我们后续的工作主要还应该进一步提高非结句点号识别的性能。

当然，由于目前的系统对其余符号的处理较为简单，所以对整体的性能会有影响，后续工作中，我们将进一步细化对其余符号的处理。

5 结论及展望

本文在基于汉语事件描述小句特点的基础上，采用了将事件描述小句识别任务转换为句中自由符号识别的方法，分析了各种类型小句的边界情况，并根据不同边界情况采用不同的分类策略。特别在非结句点号的分类中，提出了对点号类别有较大区分的两类特征，并通过实验选取合适的特征集；在后处理阶段，给出了对提高识别性能有帮助的一些规则。实验结果表明，将 EDC 识别转换为符号分类的思想是可行的。

但同时，目前的识别系统也有一些处理的盲点，如非结句点号的分类中还存在对状语层次和并列结构等情况分类效果不佳的问题。同时，对句中其余特殊符号和位置的处理过于简单，也影响了整个系统的性能。因此，后续工作中我们将以提升非结句点号分类性能为主，并辅以细化其余符号的处理方法，力求进一步提高 EDC 识别系统的识别能力。

参考文献

- [1] Steven Abney. Rapid Incremental Parsing with Repair, Proceedings of the 8th New OED Conference: Electronic Text Research, University of Waterloo, Ontario, 1990..
- [2] Wilson J. Leffa. Clause processing in complex sentences, Proceedings of LREC'98", Granada, Espanha, 1998.
- [3] Constantin Orasan. A hybrid method for clause splitting in unrestricted English texts, Proceedings of ACIDCA'2000", Monastir, Tunisia, 2000.
- [4] Sang T K and Déjean H.(2001). Introduction to the CoNLL-2001 Shared Task: Clause Identification [A]. In Proc. of CoNLL-2001 [C], Toulouse, France, p53-57.
- [5] 周强, 李玉梅.汉语块分析评测任务设计, 第十届全国计算语言学学术会议 CNCCL-2009 论文集, 北京:清华大学出版社.pp.130-135.
- [6] 宇航. 汉语句法块自动分析研究, 清华大学计算机系本科毕业设计论文, 2007 年 6 月.