基于细粒度特征的话题句识别方法[1]

蒋玉茹[1,2] 宋柔[1,3]

（北京工业大学计算机学院 北京 100022）[1]（北京信息科技大学计算机学院 北京 100101）[2]

（北京语言大学信息科学学院 北京 100083）[3]

E-mail:yurujiang@126.com,songrou@126.com

摘要：话题句识别中采用穷举方法生成标点句的候选话题句影响系统的执行效率和话题句识别的准确率，本文利用标点句在篇章中的位置特征、话题的语法特征以及话题串与说明的邻接性特征，指导候选话题句的生成过程，减少了候选话题句的个数，提高了系统效率，并使得话题句识别的准确率在现有基础上提高了 0.96 个百分点。

关键词：话题；话题句；编辑距离；语法特征；可邻接性；

中图分类号：TP391　　**文献标识码：**A

Topic Clause Identification Based On Specific Features

Jiang Yuru[1,2] Song Rou[1,3]

（Beijing University of Technology,Computer School, Beijing 100022）[1]

（Beijing Information and Science Technology University, Computer School, Beijing 100101）[2]

（Beijing Language and Culture University,Information Science School, Beijing 100083）[3]

Abstract: When identifing the topic clause (abbrieved as TC) of punctuation clause (abbrieved as PClause), the brute-force method to generate candidate topic clause (abbrieved as CTC) cause the identification system time-consuming and low accuracy. In this paper, we improve the generating process of the CTC by using specific features such as the PClause location in the text, the grammatical features of the topic and the adjacent features of topic and its comment. The experimental result shows that the improved method can not only improve the efficiency of the system by reducing the number of CTCs, but also make the recognition accuracy of TC increased 0.96 percent over the current state.

Key words: Topic；Topic Clause； Edit Distance；Grammatical Feature；Adjacent Feature；

1 引言

汉语标点句句首话题缺失是机器翻译、信息抽取准确率不高的原因之一。要解决该问题需要从汉语实际出发，研究汉语的篇章结构。

在国外，有些篇章结构理论已经开始逐步应用于自然语言处理领域，主要包括修辞结构理论（Rhetorical Structure Theory，RST）[1]和中心理论（Centering Theory，CT）[2]。RST 理论主要描述组成篇章的各个组成部分之间的修辞关系。现有研究工作包括：基于 RST 的语料标注[3]、篇章关系分析工具[4]及识别篇章结构关系[5]等。CT 理论关注篇章中指代词的形式和分布规律，及其对篇章连贯性的影响，主要用于指代消解。关于汉语篇章结构理论研究主要集中在话题[6][7][8]和回指[8][9][10][11]两个方面，还主要在语言学领域。乐明等基于 RST

[1]基金项目：国家自然科学基金（60872121，61171129，61070119）

作者简介：蒋玉茹（1978—），女，博士在读，讲师，主要研究方向为自然语言处理、语义网；宋柔（1946—），男，教授，主要研究方向为自然语言处理

理论，标注了一定规模的语料[12]。王德亮等基于 CT 理论，对汉语零形回指进行了研究[13]。计算语言学领域中，除了宋柔等[14~20]开展的广义话题理论的相关工作外，尚未见到系统性的工作。该项工作调查研究了 27 万多字，近 3 万标点句的不同语体的语料，有三分之一以上的标点句首部成分缺失而同其他标点句共享，而且由于时代、语体、作者风格的差异，有些类型的文本中这类标点句高达一半以上[16]。基于对大量语料的调查和标注工作，宋柔提出了广义话题理论[17]，该理论根据汉语篇章的特点，以边界明确的标点句为基础，提出了广义话题和话题句的概念，阐述了汉语的话题结构和话题句特征，描述了话题句动态生成的堆栈模型[16]，为汉语的篇章分析提供了理论基础和形式化的手段。

为了解决汉语标点句首话题缺失的问题，文献 19 曾经从广义话题理论出发，提出了话题句识别的研究方案：根据汉语话题结构的堆栈模型，采用穷举的策略，通过为标点句添加话题而构造候选话题句集合；采用编辑距离[21]计算候选话题句和话题句实例的相似性；并根据相似度的大小筛选候选话题句，找出正确的话题句；此外，为了克服数据稀疏，在计算编辑距离时，对候选话题句和话题句实例进行了语义泛化。在单个标点句的话题句识别的实验中，开放测试的准确率达到了 73.36%。

本文基于文献 19 的工作，在候选话题句的生成过程中，利用标点句在篇章中的位置特征、话题的语法特征，并自建本体知识库，评估话题串和说明的可邻接性，指导候选话题句的生成过程，减少了候选话题句的个数，提高了系统效率，并提高了话题句识别的准确率。

本文以下部分的组织结构是：第 2 节概述文献[19]中单个标点句话题识别过程中与本文紧密相关的内容；第 3 节提出候选话题句的生成方法及实验结果；第 4 节给出结论。

2 单个标点句话题识别[2]

2.1 概述

单个标点句话题句的识别可以描述为如下的任务：已知篇章中某个标点句 c 以及 c 的上一个标点句的话题句 t_{pre}，求标点句 c 的话题句 t。

比如在介绍鲛鳣目的文本[3]中，有：

例 1.

t_{pre}：鲛鳣目 体 无 鳞 ，

c：皮肤 裸露 或 具 硬棘 ，

如何求 c 的话题句“鲛鳣目皮肤裸露或具硬棘”？

识别过程有三个主要步骤：

步骤 1.生成候选话题句

根据堆栈模型，c 可能是完整的话题句，也可能只是说明部分，句首缺失话题串，而缺失的话题串一般在 t_{pre} 的前部。于是对于例 1 中的 c，可以生成如下候选话题句：

d_0.皮肤 裸露 或 具 硬棘 ，

d_1.鲛鳣目∽ 皮肤 裸露 或 具 硬棘 ，

d_2.鲛鳣目 体∽ 皮肤 裸露 或 具 硬棘 ，

d_3.鲛鳣目 体 无∽ 皮肤 裸露 或 具 硬棘 ，

d_4.鲛鳣目 体 无 鳞∽ 皮肤 裸露 或 具 硬棘 ，

其中“∽”表示两个词串的连接，其前部是取自于 t_{pre} 的话题串，后部是用作说明的标点句 c。

---

[2]由于篇幅所限，本文仅举例说明，详见文献[19]。

[3]本文中所有文本实例均取自《中国大百科全书》光盘版[22]

设 t_{pre}为 n 个词的串 $tw_1\cdots\cdots tw_n$，简记作 tw_1^n。它的前 i 个词组成的串 $tw_1\cdots\cdots tw_i$简记作 tw_1^i，其中 i=0…n。当 i=0 时 tw_1^i为空串。从上述候选话题句的生成过程可以看出，若 t_{pre}由 n 个词组成，则 c 的候选话题句有 n+1 个，分别由 tw_1^i后连 c 而组成，即

$d_i = tw_1^i \frown c$，其中 $i \in [0, n]$。（1）

步骤 2. 计算候选话题句的可用度

为了从候选话题句中优选出正确的话题句，需要构建训练用的话题句语料库（Topic Clause Corpus），记为 *Tcorpus*。

设 *c* 的候选话题句的集合为 *CTset*(*c*)。对 *CTset*(*c*)中每个候选话题句 *d*，找 *Tcorpus* 中与它最相似的话题句 t，它们的相似度称为 d 的可用度,记作 *sim_CT*(*d*)，可以定义为：

$$sim\_CT(d) = \max_{t \in Tcorpus} sim(d,t) \quad (2)$$

其中 *sim(d,t)*是 d 和 t 的相似度，可用编辑距离来计算：

$$sim(\mathrm{d},t) = 1 - 2 * ed(d,t)/(|d| + |t|) \quad (3)$$

其中，*ed(d, t)*代表泛化后的 d 和 t 的编辑距离。

泛化是将具体的词用其语义类别替代，泛化串中字母是语义类别的标记。如果某个词没有指定的语义类别，则泛化串中出现的是该词本身。

步骤 3. 优选候选话题句

使 *sim_CT*(*d*)最大的 *d* 就是 *c* 的优选话题句。

由于可能有不止一个候选话题句都取得最大的可用度，所以理论上说一个标点句的优选话题句会构成一个集合。

任意标点句 *c* 的优选话题句集合 *OPset*(*c*)定义为：

$$OPset(c) = \left\{ \arg\max_{d \in CTset(c)} sim\_CT(d) \right\} \quad (4)$$

其中 argmax 的功能是找出取最大值的自变量。

最后，制定某种策略（比如长度最小），从集合中选取候选话题句作为 c 的话题句。

利用上述三个步骤，文献[19]取得了 73.36%的话题句识别的准确率。

2.2 分析

从公式（2）和（3）可见，对于两个不同的候选话题句来说，在 Tcorpus 中找到各自适当的话题句后，如果两对句子的编辑距离相同，那么|d|+|t|较长的候选话题句会取得较高的可用度。然而，具有较高可用度的候选话题句却未必是正确答案。

例 2:

t_{pre}: 鲛鳣目口通常上位。

c: 上颌由前颌骨组成,

可以得到如下的候选话题句:

d_1: **鲛鳣目 ∽ 上颌 由 前颌骨 组成 ，** 泛化后为: A C 由 C 组成,

d_2: 鲛鳣目 口 ∽ 上颌 由 前颌骨 组成 ， 泛化后为: A C C 由 C 组成,

等。泛化串中，A 代表某种鱼，C 代表鱼的某个部件。

在 Tcorpus 中找到与 d_1、d_2最相似的话题句:

t_1: 剑鱼 吻 由 前颌骨 及 鼻骨 组成 ， 泛化后为: A C 由 C 及 C 组成 ，

t_2: 喉盘鱼目 背鳍 和 臀鳍 全 由 鳍条 组成 ; 泛化后为: A C 和 C 全 由 C 组

成 ；

其中，d_1与 t_1编辑距离为 2；d_2与 t_2编辑距离也为 2（注：本文所有编辑距离计算中使用的均是句子的泛化串，且不包含句末点号）。根据公式 3，因$|d_1|+|t_1|<|d_2|+|t_2|$，所以 $sim(d_1,t_1)< sim(d_2,t_2)$，最后导致 $sim\_CT(d_1)<sim\_CT(d_2)$，选 d_2为优选的话题句。

但是 d_2中“口”和“上颌”不应该邻接，正确答案应该是 d_1。从这个例子可见，基于编辑距离的方法比较粗糙，还需要改进。本文将在优选候选话题句之前，在候选话题句生成的阶段，利用细粒度的特征，消除掉一部分不可能成为答案的候选话题句，从而减少优选话题句阶段可能选错的风险。

3 应用细粒度特征筛选候选话题句

候选话题句原有的生成方法（见第 2 节公式（1））是根据堆栈模型的穷举式的方法。根据大量标注实例，一些细粒度特征可以用来避免生成不必要的候选话题句，减少后期筛选候选话题句的难度，提高话题句识别的准确率。本文使用的细粒度特征有标点句在篇章中的位置、话题的语法特征、话题串与说明的可邻接性。

3.1 标点句的位置

从对中国大百科鱼类语料的标注过程[17]中，容易看出，如果标点句处于篇首，则其话题句就是它本身。

处于非篇首的段首标点句，通常并不缺话题，如果缺话题，所缺的话题就应是作为篇章名的鱼名。

例如在介绍鲛鳚目的文本中，第二段段首标点句为：“分布 于 三 大洋 热带 及 温带 海区 。”，其首部缺少话题，所缺话题为篇名“鲛鳚目”。而在介绍澳洲肺鱼的文本中，第二段段首标点句为：“澳洲肺鱼 产卵期 很 长 ，”，其句首为“澳洲肺鱼”，不缺话题。

3.2. 话题的语法特征

3.2.1 话题的词类特征

从已经标注过的语料中，包括小说、新闻、法律等文体的语料中发现，数词、数量词、数量结构、程度副词不能充当话题[17]。

例 3：

t_{pre}： 蝙蝠鱼科 胸鳍 具 3 鳍条 基骨 ，

c： 形成 假臂 ；

如例 2 那样得到 d_i和 t_i：

d_1：蝙蝠鱼科 胸鳍 形成 假臂 ； 泛化后为：A C 形成 C ；

t_1：银汉鱼目 腹鳍 不 形成 抱交器 ， 泛化后为：A C 不 形成 C ，

d_2：蝙蝠鱼科 胸鳍 具 3 形成 假臂 ； 泛化后为：A C 具 s 形成 C ；

t_2：帆蜥鱼 腹鳍 具 9 鳍条 ； 泛化后为：A C 具 s C ；

等。其泛化串中 s 是数字的泛化符号。

例 3 中，“形成假臂”的应该是“胸鳍”，即 c 的话题句应该为 d_1。但实际优选出的话题句是 d_2，其中数词 3 充当了话题，是错误的。

3.2.2. 话题的短语特征

标点句中如果一些词被“和”、“与”、“及”、“同”、“或”、“、”连接起来，且这些词为同一个泛化类别，这样的词串可称为并列型的短语结构。我们发现，并列型的短语结构，其整体可以作为话题，但是其部分片段不能成为话题。

例 4：

t_{pre}： 鲛鳐目 一般 均 具 瓣膜状 或 球茎状 吻触手 ，

c： 用以 引诱 食饵 。

与上例一样得到 d_i 和 t_i：

d_1： 鲛鳐目 一般 均 具 **瓣膜状** 用以 引诱 食饵 。

泛化后为：A 一般 均 具 H 用以 引诱 食饵 。

t_1： 喉盘鱼目 卵 一般 具 粘性 ，

泛化后为：A C 一般 具 H ，

d_2： 鲛鳐目 一般 均 具 **瓣膜状 或** 用以 引诱 食饵 。

泛化后为：A 一般 均 具 H 或 用以 引诱 食饵 。

t_2： 同 t_1

d_3： 鲛鳐目 一般 均 具 **瓣膜状 或 球茎状** 吻触手 用以 引诱 食饵 。

泛化后为：A 一般 均 具 H 或 H C 用以 引诱 食饵 。

t_3： 绯鱵 两 颌 均 具 绒毛状 牙带 ；

泛化后为：A s C 均 具 H C ；

候选话题句 d_1 和 d_2 的可用度比 d_3 的可用度高，但是 d_1 中的“瓣膜状”和 d_2 中的“瓣膜状 或”都是并列结构“瓣膜状 或 球茎状”的一部分，不应该成为话题，d_3 才是正确的话题句。

3.3. 话题串和说明的可邻接性

话题串 tw_T^i 和标点句 c 都是合法的标点句的片段，它们内部的词序列是正确的，但这两个串连成的新串却不一定合法，原因是在连接后，tw_T^i 和 c 邻接处的句法和语义关系中存在非法的情况。如例 2 中“口”和“上颌”。

话题串与说明的可邻接性定义为：话题串 tw_T^i 的最后一个成分 tw_T 和标点句 c 的第一个成分 cw_1 是否可以邻接共现。

在鱼类百科文本的行文中，如果一个邻接词对均为鱼名，则它们具有上下位关系；如果一个邻接词对均为部件，或前者是鱼名，后者是部件，则它们具有整体部件关系。

据此，可构造一个本体知识库 OntoCorpus，其中包含词对<a,b>，当且仅当 a、b 在语料库中邻接共现，而且满足下列条件之一：

- gen(a)=A 且 gen(b)=A
- gen(a)=C 且 gen(b)=C
- gen(a)=A 且 gen(b)=C

其中，A 代表鱼名，C 代表部件，gen(w)代表 w 的语义类别。

利用 OntoCorpus，在生成候选话题句时，可以判断鱼名和鱼名、鱼名和部件、部件和部件的可邻接性，将出现不可邻接情况的候选话题句预先筛掉。

3.4. 基于细粒度特征的候选话题句生成算法

根据前面的分析，可以制定如下的候选话题句生成方法：

[1].处于篇首的标点句，其候选话题句就是它本身；

[2].处于非篇首的段首标点句，其候选话题句生成策略如下：

如果它的首词是鱼名，则它的候选话题句就是它本身；

如果它的首词是鱼的部件名，那么它应是缺少话题的标点句，而且所缺的话题就是它所在篇章的篇名（是一个鱼名）。因此，它的候选话题句就是“篇名∽标点句”。

否则，它的候选话题句有两个，一个就是它本身，另一个是“篇名∽标点句”。

[3].处于其他位置的标点句需要如本文第 2 节的步骤 1 所述穷举策略，并利用 3.2 节和 3.3 节的约束条件生成候选话题句。算法如下：

已知篇章中某个标点句 c 和它的上一个标点句的话题句 t_{pre}，

设 t_{pre} 为 n 个词的串 $tw_1\cdots\cdots tw_n$，简记作 tw_1^n，它的前 i 个词组成的串 $tw_1\cdots\cdots tw_i$ 简记作 tw_1^i。c 的第一个词记作 cw_1。

```
for k←1 to n
    if ((twk是数词、数量词、数量结构、程度副词之一)
        or( k<n and ( twk或 twk+1∈{“和”、“与”、“及”、“同”、“或”、“、”}))
        or(gen(twk)=A and gen(cw1)=A) and <twk,cw1>∉OntoCorpus)
        or(gen(twk)=C and gen(cw1)=C) and <twk,cw1>∉OntoCorpus)
        or(gen(twk)=A and gen(cw1)=C) and <twk,cw1>∉OntoCorpus))
    then 执行下一轮循环
    else 将 tw1k∽c 加入 c 的候选话题句列表并执行下一轮循环
```

以上算法生成了标点句的候选话题句集合，然后再利用公式 2~4 做进一步优选，以识别出话题句。

3.5. 实验及结果

3.5.1 语料

为了使本文的实验结果与文献[19]的实验结果对照，采用了与它相同的实验数据。即从中国大百科全书生物卷中选取关于鱼的文本 202 篇，其中包含 9508 个标点句。测试语料为其中 15 篇文本，共 717 个标点句，训练语料为其余的 187 篇文本。OntoCorpus 中本体词对从所有 202 篇文本的话题句中提取。

对 9508 个标点句对应的话题句进行统计，话题串中词数的分布如图 2：

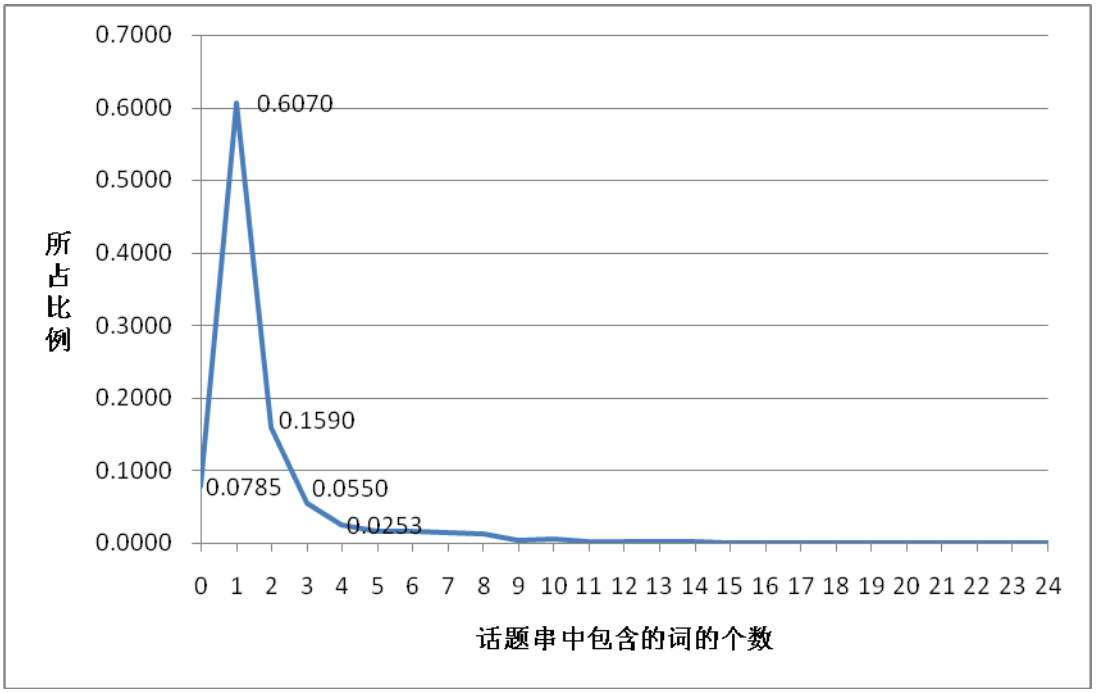

图 2 话题串包含词数的分布图

话题串包含 1 个词的话题句所占比例最大，为 60.70%；包含 2 个词的话题句所占比例为 15.90%；话题串为空的占 7.85%；话题串长度在 5 个词以上的占 7.52%，最长的话题串中包含了 24 个词。

3.5.2 结果

文献[19]的实验过程为：步骤 1：构造候选话题句→步骤 2：计算候选话题句的可用度→步骤 3：优选候选话题句。

本文对“步骤 1：构造候选话题句”这一阶段进行了优化。产生了两个效果：

[1].提高系统效率。

针对测试语料中的 717 个标点句，如果采用第 2 节的穷举式的候选话题句生成策略（即文献[19]中的构造候选话题句的方法），可生成 5118 个候选话题句，而采用 3.4 节的方法之后，候选话题句的数量降到 4169 个，减少了 18.54%，提高了系统执行效率。

[2].提高话题句识别的准确率。

图 3 中展示了不同候选话题句筛选策略对话题识别准确率的影响，其中 test1 为穷举式的候选话题句生成的实验结果，即文献[19]的实验结果，准确率为 73.36%；test2 为在 test1 基础上考虑标点句位置的实验结果，准确率为 73.64%；test3 为在 test2 基础上考虑话题的语法特征的实验结果，准确率为 74.20%；test4 为在 test3 基础上考虑话题和说明的可邻接性的实验结果，准确率为 76.15%，可以看出利用本体知识库评估话题和说明的可邻接性对话题句识别的准确率可起到明显的结果，其他两种策略也有一定作用。

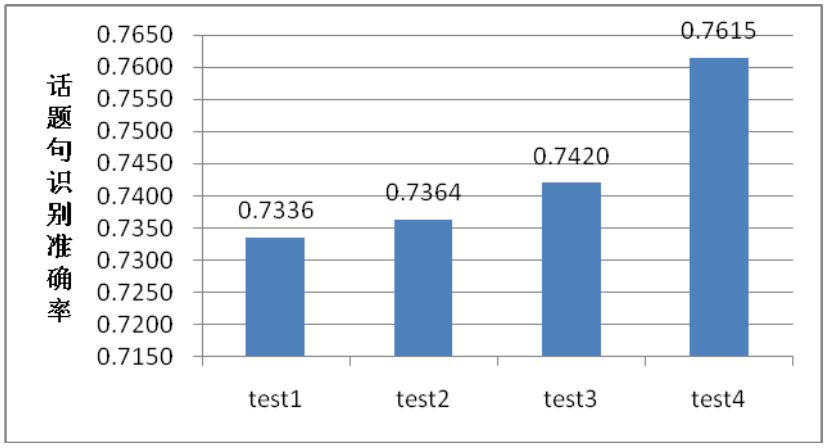

图 3 候选话题句筛选策略与话题句识别准确率关系

文献[20]对文献[19]的步骤 2 所采用的评估函数进行了优化，将候选话题句可用度的评估函数分为：（1）基于整句相似性的评估函数；（2）基于上下文相似性的评估函数；（3）整句相似性与局部相似性结合的评估函数；（4）上下文相似性与局部相似性结合的评估函数。其中第四个评估函数取得的实验效果最佳。

从中国大百科全书生物卷中选取关于鱼的文本 200 篇，分成 10 份，进行十折交叉验证[23]实验，结果见图 4。其中方案 A 采用的是文献[19]的实验过程；方案 B 采用的是文献[20]的实验过程（第四个评估函数）；方案 A+C 和方案 B+C 是分别将方案 A 和 B 的候选话题句生成算法替换为本文提出的算法。

从图 3 可以看出，方案 A+C 的十折交叉验证的均值为 71.28%，比方案 A 提高了 1.89 个百分点；方案 B+C 的均值为 77.24%，比方案 B 提高了 0.96 个百分点。

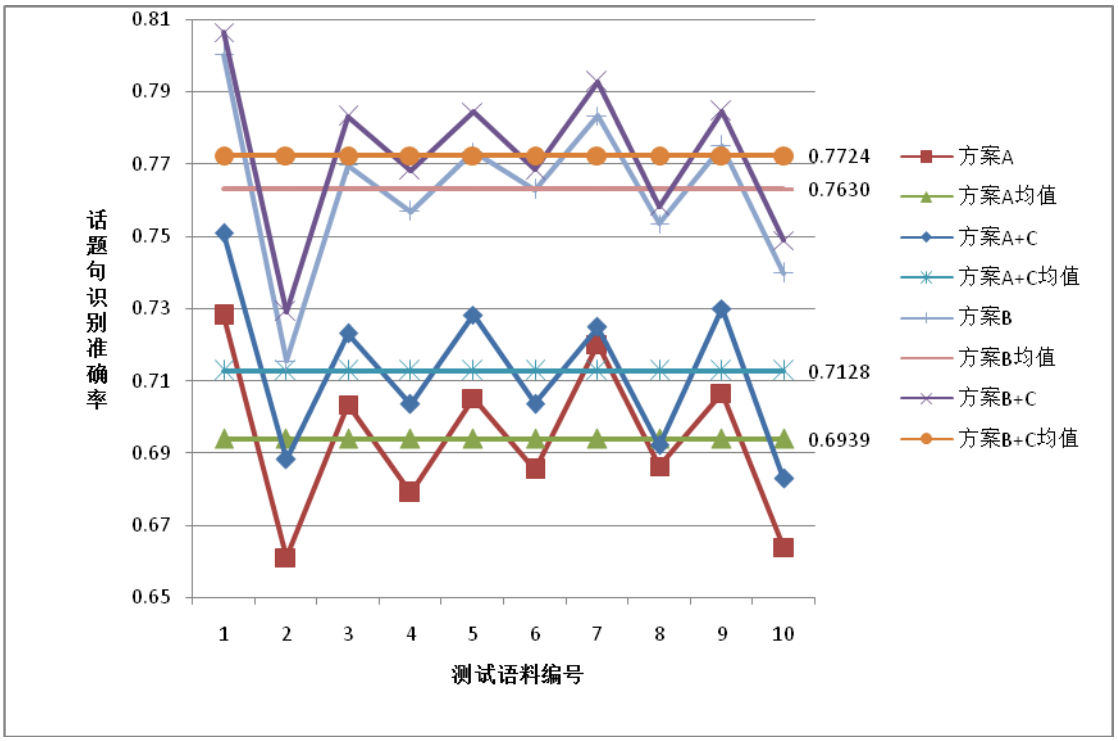

图 4 话题句识别的十折交叉验证结果

3.6 分析

当前采用的话题句生成方法基于这样的基本假设：标点句的候选话题句可以用堆栈模型

生成，具体说来即（1）它本身一定是完整的说明，（2）它的话题一定是它前一个标点句的话题句的某一个前缀部分。但是存在违背这一假设的情况：

[1].标点句后部不完整。

例 5:

t_{pre}: 鼹鱼亚目 **为 一 群** 身体 较高 ，

c_1: 具 鲜艳 条纹 或 斑纹 ，

c_2: 生活 于 热带 海洋 **的 小鱼** ，

c_1 的候选话题句有：

d_1: 鼹鱼亚目 具 鲜艳 条纹 或 斑纹 ，

d_2: 鼹鱼亚目 为 一 群 具 鲜艳 条纹 或 斑纹 ，

等。

例 5 中“为一群……的小鱼”是前后呼应的结构，但是它们没有出现在同一个标点句中，跨越了标点句的范围，导致 c_1 的话题句 d_2 缺少了“的小鱼”这个部分，不是完整的说明，于是实验系统错误地选择了 d_1。

一般而言，这种跨越标点句的搭配结构无法用堆栈模型处理，而需要用汇流模型[18]。如果能够正确的识别跨越标点句的搭配结构，可以进一步减少候选话题句的生成数量，提高话题句识别的正确率。

[2].标点句说明的话题在上一个标点句的话题句的后部。

例 6：

t_{pre}: 中国 玻甲鱼科 另 一 种 为 条纹虾鱼 ，

c: 眼间 凸 而 不 凹 ，

根据堆栈模型，它的候选话题句有：

d_1: 中国 玻甲鱼科 眼间 凸 而 不 凹 ，

d_2: 中国 玻甲鱼科 另 一 种 为 条纹虾鱼 眼间 凸 而 不 凹 ，

等。

因为 c 是对“条纹虾鱼”的说明，c 的话题句应该是

条纹虾鱼 眼间 凸 而 不 凹 ，

依据堆栈模型的穷举方法不可能生成出这样的话题句。如要生成这样的话题句，需要采用节栈模型[18]。但是如何识别出标点句 c 的说明范围是话题串的整体还是局部，这一问题有待研究。

4 总结

在候选话题句的生成过程中，利用标点句在篇章中的位置、话题的语法特征、话题串和评述的邻接性这三个细粒度特征，指导候选话题句的生成过程，能够有效减少不必要的候选话题句的生成，提高系统的执行效率，并提高话题句识别的准确率。但是在识别跨越标点句范围的搭配结构，识别标点句说明的对象的边界方面还需要做进一步的工作。

参考文献

[1] Mann. W. & S. Thompson. Rhetorical Structure Theory :A Theory of Text Organization[R]. US:USC Information Science Institute. Technical Report I (SI/ RS - 87 - 190) ,1987.

[2] GROSZ, BARBARA J., ARAVIND K. JOSHI, AND SCOTT WEINSTEIN.CENTERING: A FRAMEWORK

FOR MODELING THE LOCAL COHERENCE OF DISCOURSE[J]. COMPUTATIONAL LINGUISTICS,1995, 21 (2): 203-225.
[3] Carlson ,L. ,Marcu. D. & Okurowski M. Building a Discourse tagged Corpus in the Framework of Rhetorical Structure Theory[C]. Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics ,Seattle ,WA ,2001 :9 - 17.
[4] HILDA: A Discourse Parser Using Support Vector Machine Classification[J]，Dialogue and Discourse 1(3) ,2010: 1-33
[5] An Unsupervised Approach to Recognizing Discourse Relations[C]，Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002: 368-375.
[6] 曹逢甫.汉语的句子与子句结构[M].北京：北京语言大学出版社，2005 年：19~56
[7] 屈承熹.汉语篇章语法[M]. 北京：北京语言大学出版社，2006 年：191~292
[8] 徐赳赳.现代汉语篇章语言[M]学.北京：商务印书馆，2010 年：287~363
[9] 陈平.汉语零形回指的话语分析[J].中国语文.1987 年，5：363~378
[10] 徐赳赳.现代汉语篇章回指研究[M].北京：中国社会科学出版社，2003 年：12~270
[11] 许余龙.篇章回指的功能语用探索[M].上海：上海外语教育出版社，2004 年：179~248
[12] 乐明.汉语篇章修辞结构的标注研究[J].中文信息学报，2008 年 7 月，22（4）：19~23
[13] 王德亮. 汉语零形回指解析_基于向心理论的研究[J].现代外语（季刊）.2004 年 11 月，27（4）：350~359
[14] 张瑞朋.现代汉语书面语中跨标点句句法关系约束条件的研究[D].北京语言大学博士论文.2007
[15] 黄健传，宋柔.标点句标注研究[C].//孙茂松，陈群秀.内容计算的研究与应用前沿（第九届全国计算语言学学术会议论文集）.北京：清华大学出版社，2007
[16] 宋柔，现代汉语跨标点句句法关系的性质研究[J]. 世界汉语教学，2008，2,26~44
[17] Rou Song, Yuru Jiang, Jingyi Wang. On Generalized-Topic-Based Chinese Discourse Structure[C]. Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing：清华大学出版社，2010, 23-33.
[18] 宋柔，汉语篇章广义话题结构研究[R]，北京：北京语言大学语言信息处理研究所，2012 年 5 月
[19] 蒋玉茹，宋柔. 基于广义话题理论的话题句识别[J].中文信息学报，2012，26(5),114~119
[20] 蒋玉茹，宋柔.话题句识别中候选话题句评估函数的优化. 已投稿.2013
[21] Michael Gilleland, Levenshtein Distance, in Three Flavors[OL], http://www.merriampark.com/ld.htm
[22] 中国大百科全书出版社.中国大百科全书光盘版[CD].1999. 中国大百科全书出版社
[23] Ron Kohavi.A study of cross-validation and bootstrap for accuracy estimation and model selection[A]. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2[C]. San Mateo: Morgan Kaufmann, 1995，1137–1143.