

基于细粒度特征的话题句识别方法

蒋玉茹^{1,2*}, 宋 柔^{1,3}

(1. 北京工业大学 计算机学院, 北京 100022; 2. 北京信息科技大学 计算机学院, 北京 100101;

3. 北京语言大学 信息科学学院, 北京 100083)

(* 通信作者电子邮箱 yurujiang@126.com)

摘 要: 话题句(TC)识别中采用穷举方法生成标点句的候选话题句(CTC)影响系统的执行效率和话题句识别的准确率。提出一种新的候选话题句生成方法,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。实验结果表明,该方法减少了候选话题句的个数,提高了系统效率。而且,通过与基于穷举式候选话题句生成策略的话题句识别工作进行对比,该方法使单个标点句话题句识别的准确率提高了0.96个百分点,使标点句序列话题句识别的准确率提高了1.31个百分点。

关键词: 话题句; 编辑距离; 语法特征; 可邻接性

中图分类号: TP391 **文献标志码:** A

Topic clause identification method based on specific features

JIANG Yuru^{1,2*}, SONG Rou^{1,3}

(1. College of Computer Science, Beijing University of Technology, Beijing 100022, China;

2. Computer School, Beijing Information Science and Technology University, Beijing 100101, China;

3. College of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: When identifying the Topic Clause (TC) of Punctuation Clause (PClause), the brute-force method to generate Candidate Topic Clause (CTC) causes high time consumption and low accuracy of the identification system. A new CTC generating method was proposed, which used 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 accuracy of TC identification for single PClause and PClause sequence increase by 0.96 percentage points and 1.31 percentage points respectively over the current state.

Key words: Topic Clause (TC); edit distance; grammatical feature; adjacent feature

0 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。

计算语言学领域中,除了文献[1-7]开展的广义话题理论的相关工作外,尚未见到系统性的关于汉语篇章结构的研究工作。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句首成分缺失而与其他标点句共享,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^[3]。基于对大量语料的调查和标注工作, Song等提出了广义话题理论^[4],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^[3],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句首话题缺失的问题,文献[6]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^[8]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73.36%。

本文基于文献[6]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。

1 单个标点句话题识别

1.1 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_c 。

收稿日期:2013-10-29;修回日期:2014-01-08。

基金项目:国家自然科学基金资助项目(60872121, 61171129);北京市教委科研计划面上项目(KM201411232014)。

作者简介:蒋玉茹(1978-),女,辽宁沈阳人,讲师,博士研究生,CCF会员,主要研究方向:自然语言处理; 宋柔(1946-),男,江苏苏州人,教授,CCF会员,主要研究方向:自然语言处理。

比如在介绍鮫鰈目的文本(本文所有示例均引自文献[9])中,有:

例1

t_{pre} : 鮫鰈目 体 无 鳞 ,

c : 皮肤 裸露 或 具 硬棘 ,

如何求 c 的话题句“鮫鰈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 : 鮫鰈目 \sim 皮肤 裸露 或 具 硬棘 ,

d_2 : 鮫鰈目 体 \sim 皮肤 裸露 或 具 硬棘 ,

d_3 : 鮫鰈目 体 无 \sim 皮肤 裸露 或 具 硬棘 ,

d_4 : 鮫鰈目 体 无 鳞 \sim 皮肤 裸露 或 具 硬棘 ,

其中“ \sim ”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 。

设 t_{pre} 为 n 个词的串 $tw_1 \cdots tw_n$, 简记作 tw_1^n 。它的前 i 个词组成的串 $tw_1 \cdots tw_i$ 简记作 tw_1^i , 其中 $i = 0, 1, \dots, n$ 。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:

$$d_i = tw_1^i \sim c; i \in [0, n] \quad (1)$$

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 $score(d)$,可以定义为:

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

其中: $sim(d, t)$ 是 d 和 t 的相似度,可用编辑距离来计算:

$$sim(d, t) = 1 - 2 \times \frac{ed(d, t)}{|d| + |t|} \quad (3)$$

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 $score(d)$ 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:

$$OPset(c) = \{ \arg \max_{d \in CTset(c)} score(d) \} \quad (4)$$

其中: $\arg \max$ 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[6]取得了73.36%的话题句识别的准确率。

1.2 分析

从式(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} : 鮫鰈目 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(泛化串中, A 代表某种鱼, C 代表鱼的某个部件。):

d_1 : 鮫鰈目 \sim 上颌 由 前颌骨 组成,

泛化后为: $A C$ 由 C 组成,

d_2 : 鮫鰈目 口 \sim 上颌 由 前颌骨 组成,

泛化后为: $A C C$ 由 C 组成,

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 $score(d_1) < score(d_2)$, 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。

2 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。

2.1 标点句的位置

从对中国大百科鱼类语料的标注过程^[4]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。

例如在介绍鮫鰈目的文本中,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“鮫鰈目”。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。

2.2 话题的语法特征

2.2.1 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^[4]。

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_1 、 t_1 、 d_2 和 t_2 等(泛化串中 s 是数字的泛化符号):

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 $s C$;

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。

2.2.2 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。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} :鲛鰈目 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 :鲛鰈目 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 :鲛鰈目 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 :喉盘鱼目 卵 一般 具 粘性,

d_3 :鲛鰈目 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 :绯纈 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。

2.3 话题串和说明的可邻接性

话题串 tw_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_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_i 的最后一个成分 tw_i 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 $\langle a, b \rangle$,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:

1) $gen(a) = A$ 且 $gen(b) = A$;

2) $gen(a) = C$ 且 $gen(b) = C$;

3) $gen(a) = A$ 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。

2.4 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 \wedge 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 \wedge 标点句”。

3)处于其他位置的标点句需要利用本文第1.1节的步骤1所述穷举策略及2.2节和2.3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, $tw_1 \dots tw_n$,简记作 tw_n^i ,它的前 i 个词组成的串 $tw_1 \dots tw_i$ 简记作 tw_i^i 。 c 的第一个词记作 cw_1 。

for $k \leftarrow 1$ to $n + 1$

if ($(tw_k$ 是数词、数量词、数量结构、程度副词之一)

or ($k < n$ and (tw_k 或 $tw_{k+1} \in \{“和”、“与”、“及”、“同”、“或”、“、”\}$)) or

($gen(tw_k) = A$ and $gen(cw_1) = A$) and

$\langle tw_k, cw_1 \rangle \notin OntoCorpus$) or

($gen(tw_k) = C$ and $gen(cw_1) = C$) and

$\langle tw_k, cw_1 \rangle \notin OntoCorpus$) or

($gen(tw_k) = A$ and $gen(cw_1) = C$) and

$\langle tw_k, cw_1 \rangle \notin OntoCorpus$))

then 执行下一轮循环

else 将 $tw_i^k \wedge c$ 加入 c 的候选话题句列表并执行下一轮循环

以上三步生成了标点句的候选话题句集合,然后再利用式(2)~(4)作进一步优选,以识别出话题句。

2.5 实验及结果

2.5.1 开放测试结果

本文从中国大百科全书生物卷中选取关于鱼的文本243篇,其中包含11790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228篇文本。 $OntoCorpus$ 中本体词对从所有243篇文本的话题句中提取。

文献[6]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1.1节中步骤1所述的穷举式的候选话题句生成策略(即文献[6]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2.4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1中展示了不同候选话题句筛选策略对话题句识别准确率的影响,其中:test1为穷

举式的候选话题句生成的实验结果,准确率为73.64%;test2为在test1基础上考虑标点句位置的实验结果,准确率为74.48%;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74.76%;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75.59%,可以看出三种筛选策略对话题句识别的准确率均可起到积极的作用。从筛选策略的构造过程也可以看出,它不会筛选掉正确的话题句,因此对话题句的识别没有负面的影响。

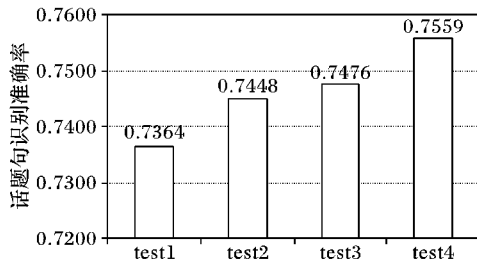


图1 候选话题句筛选策略与话题句识别准确率关系

2.5.2 十折交叉验证测试结果

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^[10]测试实验。

单个标点句话题句识别的十折交叉验证测试结果如下:

文献[7]对文献[6]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。

图2中方案A采用的是文献[6]的实验过程;方案B采用的是文献[7]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。

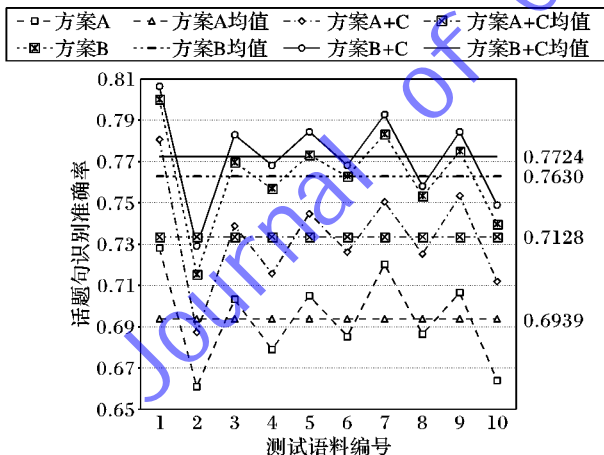


图2 单个标点句话题句识别的十折交叉验证测试结果

从图2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。说明本文所提出的候选话题句生成方法对单个标点句的话题句识别有一定积极作用。

标点句序列话题句识别的十折交叉验证测试结果如下:

文献[11]是对标点句序列的话题句识别工作,在其基础上应用基于上下文可用度和局部可用度的评估函数^[7],即图

3中的方案A;在方案A的基础上应用本文方法,即图3中的方案B,图3中展示了方案A和方案B的十折交叉验证测试结果。

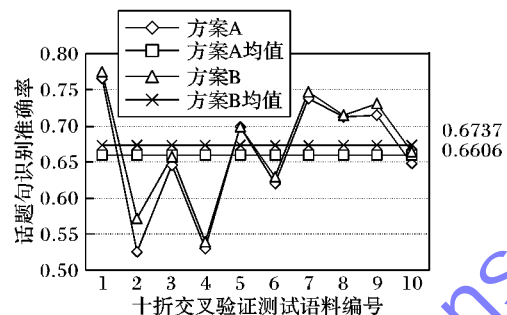


图3 标点句序列话题句识别的十折交叉验证测试结果

从图3可以看出,方案A的话题句识别准确率均值为66.06%,方案B的话题句识别准确率为67.37%,比方案A提高了1.31个百分点。说明本文所提出的候选话题句生成方法对标点句序列的话题句识别也有促进作用。

2.6 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,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} : 鳃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 : 鳃鱼亚目具鲜艳条纹或斑纹,

d_2 : 鳃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^[5]。如果能够正确地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^[5]。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。

3 结语

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,并提高单个标点句话题句识别和标点句序列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,识别标点句的说明对象的边界方面还需要做进一步的工作。

参考文献:

- [1] ZHANG R. The research about the constraint rules of syntax relation in cross-punctuation sentence in written modern Chinese [D]. Beijing: Beijing Language and Culture University, 2007. (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: 北京语言大学, 2007.)
 - [2] HUANG J, SONG R. A research on punctuation annotation [C]// Proceedings of the 9th China National Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2007: 350 - 355. (黄健传, 宋柔. 标点句标注研究[C]// 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: 350 - 355.)
 - [3] SONG R. A research on the properties of syntactic relation between P-clauses in modern Chinese [J]. Chinese Teaching in the World, 2008, 2: 26 - 44. (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.)
 - [4] SONG R, JIANG Y, WANG J. On generalized-topic-based Chinese discourse structure [C]// SIGHAN 2010: Proceedings of CIPS - SIGHAN Joint Conference on Chinese Language Processing. Beijing: Tsinghua University Press, 2010: 23 - 33.
 - [5] SONG R. A research on generalized topic structure of Chinese text [R]. Beijing: Beijing Language and Culture University, 2012. (宋柔. 汉语篇章广义话题结构研究[R]. 北京: 北京语言大学, 2012.)
 - [6] JIANG Y, SONG R. Topic clause identification based on generalized topic theory [J]. Journal of Chinese Information Processing, 2012, 26(5): 114 - 119. (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.)
 - [7] JIANG Y, SONG R. Optimization of candidate topic clause evaluation function in topic clause identification [J]. Journal of Beijing University of Technology, 2014, 40(1): 43 - 48. (蒋玉茹, 宋柔. 话题句识别中候选话题句评估函数的优化[J]. 北京工业大学学报, 2014, 40(1): 43 - 48.)
 - [8] GILLELAND M. Levenshtein distance, in three flavors [EB/OL]. [2013-02-04]. <http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm>.
 - [9] HU Q. Encyclopedia of China [M/CD]. Beijing: Encyclopedia of China Publishing House, 1999. (胡乔木. 中国大百科全书: 图文数据光盘[M/CD]. 北京: 中国大百科全书出版社, 1999.)
 - [10] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]// IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1995, 2: 1137 - 1143.
 - [11] JIANG Y, SONG R. Topic structure identification of PClause sequence based on generalized topic theory [C]// Proceedings of the 2012 1st CCF Conference on Natural Language Processing and Chinese Computing. Berlin: Springer-Verlag, 2012: 85 - 96.
-
- (上接第1321页)
- [5] KHAZAEI S, HASANZADEH M M, KIAEI M S. Linear sequential circuit approximation of grain and trivium stream ciphers [EB/OL]. [2013-08-12]. <http://eprint.iacr.org/2006/141.pdf>.
 - [6] FISCHER W, GAMMEL B M, KNIFFLER O, et al. Differential power analysis of stream ciphers [C]// Proceedings of the 2007 Cryptographers' Track at the RSA Conference. Berlin: Springer-Verlag, 2007, 4377: 257 - 270.
 - [7] YUSEOP L, KITAE J, JAECHUL S, et al. Related-key chosen IV attacks on Grain-v1 and Grain-128 [C]// Proceedings of the 13th 2008 Australasian Conference. Berlin: Springer-Verlag, 2008, 5107: 321 - 335.
 - [8] DINUR I, SHAMIR A. Breaking Grain-128 with dynamic cube attacks [C]// Proceedings of the 18th International Workshop on Fast Software Encryption. Berlin: Springer-Verlag, 2011: 167 - 187.
 - [9] SONG H, FAN X, WU C, et al. Cube attack on Grain [J]. Journal of Software, 2012, 23(1): 171 - 176. (宋海欣, 范修斌, 武传坤, 等. 流密码算法 Grain 的立方攻击[J]. 软件学报, 2012, 23(1): 171 - 176.)
 - [10] BERZATI A, CANOVAS C, CASTAGNOS G, et al. Fault analysis of Grain-128 [C]// Proceedings of the 2009 IEEE International Workshop on Hardware-oriented Security and Trust. Piscataway: IEEE Press, 2009: 7 - 14.
 - [11] KARMAKAR S, CHOWDHURY D R. Fault analysis of Grain-128 by targeting NFSR [C]// AFRICACRYPT '11: Proceedings of the 4th International Conference on Progress in Cryptology in Africa. Berlin: Springer-Verlag, 2011: 298 - 315.
 - [12] LANO J, MENTENS N, PRENEEL B, et al. Power analysis of synchronous stream ciphers with resynchronization mechanism [EB/OL]. [2013-08-17]. <http://www.cosic.esat.kuleuven.be/publications/article498.pdf>.
 - [13] MANGARD S, OSWALD E, POPP T. Power analysis attacks [C]// FENG D, ZHOU Y, LIU J, et al. translation. Beijing: Science Press, 2010: 97 - 101. (MANGARD S, OSWALD E, POPP T. 能量分析攻击[M]. 冯登周, 周永彬, 刘继业, 等译. 北京: 科学出版社, 2010: 97 - 101.)
 - [14] LI L, LI R, TONG Y, et al. Development on power analysis attack and defense of embedded cipher chip [J]. Journal of Computer Research and Development, 2010, 47(4): 595 - 604. (李浪, 李仁发, 童元满, 等. 嵌入式加密芯片功耗分析攻击与防御研究进展[J]. 计算机研究与发展, 2010, 47(4): 595 - 604.)
 - [15] LIU H, ZHAO X, WANG T, et al. Research on Hamming weight-based algebraic side-channel attacks on SMS4 [J]. Chinese Journal of Computers, 2013, 36(6): 1183 - 1193. (刘会英, 赵新杰, 王韬, 等. 基于汉明重的 SMS4 密码代数旁路攻击研究[J]. 计算机学报, 2013, 36(6): 1183 - 1193.)