文章编号：1003-0077（2011）00-0000-00

基于最大熵模型的汉语标点句缺失话题自动识别初探*

卢达威[1]，宋柔[2]

（1.北京大学中文系，北京市 100871；2. 北京语言大学语言信息处理研究所，北京市 100083）

摘要：本文以判别标点句缺失话题是上句的主语还是宾语为任务，将该任务作为标点句缺失话题自动识别研究的切入点。本文首先归纳了一系列判别这一任务的字面特征和语义特征，然后结合规则和最大熵模型，进行自动判别实验。结果显示，对特定类别动词的实验 F 值达到 82%。对实验结果的分析说明，动词特征和语义特征对判别该任务的作用最大，规则方法和统计方法在判别任务中不能偏废，精细化的知识对判别的性能有重要影响。

关键字：广义话题结构；新支话题；自动识别；最大熵模型

中图分类号：TP391　　**文献标识码：**A

Study on Automatic Recognition of the Absent Topic in Chinese Punctuation Clause Based on Maximum Entropy Model

LU Dawei[1], SONG Rou[2]

(1. Peking University ,Beijing, 100871; 2. Beijing Language and Culture University, Beijing 100083)

Abstract：This paper focuses on the task of the automatic recognition, which is whether the absent topic of punctuation clause is the subject or object of the previous sentence. We regard this task as the pointcut of the automatic recognition of absent topic in Chinese punctuation clause. Several literal features and semantic features are raised to achieve this task by combining the rules and the maximum entropy model. Experimental results show that F-score of this recognition reaches 82% for the samples of specific verbs. The analysis for the experimental results shows that verbs features and semantic features play the most important role in recognition process, neither rules nor statistics can be neglected, and knowledge of refinement has great influence on the performance of the recognition .

Keywords: Generalized Topic Structure;New Branch Topic; Automatic recognition; Maximum Entropy Model

1. 引言

标点句是指汉语文本中逗号、分号、句号、叹号、问号、直接引语的引号以及这种引号前的冒号所分隔出的词语串，是汉语篇章的基本单位[1][2][3]。文献[4]在大规模语料库统计中发现，汉语篇章中，标点句的话题缺失是常态。如：

例 1：c1:三人到汽车站“留言板”上看见李顾留的纸条，

c2: ‖ 说住在火车站旁一家旅馆内，

c3: 便搬去了。

上例有 3 个标点句，除 c1 的话题—说明结构完整外，c2、c3 都缺话题。标点句的话题缺失对机器翻译、文本摘要等都是一大挑战。话题属于语用范畴的问题，不容易通过统计获得。

细读 c2、c3 发现，它们所缺话题并不一样。c2 的话题是“李顾留的纸条”，是 c1 的宾语；c3 的话题是“三人”，是 c1 的主语。在英语中，话题是上句的宾语还是主语可以用一定的形式手段来表达，如 c2 可以用为关系从句来表现。汉语缺乏形式标记，虽然汉语母语

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61171129），国家重点基础研究计划（973 计划）项目（2014CB340502）。

者也很容易凭借语感判断标点句所缺的话题是什么，但让计算机自动判别就十分困难。另外，通过对大规模语料的调查发现，标点句所缺话题除了上句的主语和动词宾语外，还可能是上句的介词宾语、主谓谓语句小主语、从句主语等，甚至上句整句作为话题[5]。因此，计算机自动识别标点句缺失话题是一项十分困难的任务。

文献[6]针对百科全书语料通过人工语义泛化标注和计算相似度的方法来识别标点句的缺失话题，F 值达到 73.64%，文献[7][8]又作了改进。由于上述方法需要大量语料标注，且针对百科全书语料的，对通用语料来说，全面的语义泛化十分困难。本文尝试从另一个角度切入，将问题的范围限定为：仅对给定的样本，区分标点句所缺话题是上句的主语还是宾语。这样就将问题简化为样本的二值分类问题。本文首先从语言学和认知方面入手，挖掘上下文的特征，进而使用统计机器学习方法，学习各个特征的权重，实现计算机的自动判别。

2. 实验样本选择

2.1. 样本定义

（1）每个样本以文本中相邻的两个标点句为原型，由前一标点句的话题自足句[1]和后一标点句组成一个句对。其中，前一个标点句的话题自足句称为**上句**，后一个标点句称为**本句**。

（2）上句必须是主动宾结构；

（3）本句必须是缺话题的标点句，而且话题一定在上句出现，且话题不是上句主语就是上句宾语。

如果样本中本句所缺话题是上句的宾语，这类样本称为**新支样本**，本句称为**新支句**，上句的宾语称为**新支话题**，如例 1 中 c1 的“李顾留的纸条”就是新支话题；如果本句所缺话题是上句的主语，则该样本称为**非新支样本**。

2.2. 实验目标

本文实验的目标就是让计算机自动区分新支样本和非新支样本。

2.3. 新支样本的情况

本文以“广义话题结构标注语料”[2]为基础，从中筛选出所有符合以上要求的新支样本，共有 431 例。另从《围城》全文和北京语言大学 CCRL 语料库的部分语料中，抽取符合要求的新支样本 228 例。实验的新支样本合共 641 例，组成新支样本集。

经统计，在新支样本集中，引出新支话题的不同的动词（如例 1 的“看见”）共有 267 个，这些动词的词形在本文实验中将直接作为实验的特征。

2.4. 非新支样本的情况

我们以这 267 个动词为基础来筛选非新支样本，筛选条件是：非新支样本的上句必须主动宾齐全，且动词必须是这 267 个词，同时本句以上句主语为话题。在“广义话题结构标注语料”中，总选出符合上述条件的非新支样本集 1508 例。

3. 特征的分析与标注

通过对新支样本和非新支样本的详细分析发现，影响本句话题所指的上下文因素非常复杂，涉及句法、语义、语用、常识，甚至专业知识等。从工程计算的角度，我们将这些特征分为六类：动词特征、接续特征、信息量特征、句法特征、语义特征和其他特征。

3.1. 动词特征

动词特征是指以样本上句主动宾句式中动词的词形为特征（即 2.3 节中提到的 267 个动词）。文献[10]对动词引出新支话题能力已有初步研究。由于每个动词对于是否能带宾语，以及所带宾语是否容易作为下一个标点句的话题，能力不一样，因此动词词形本身具有重要

[1] 标点句如果不缺少话题或说明，则本身就是话题自足句，否则按照广义话题结构流水模型的规律，从上下文补足所缺的话题或说明，补足后称为话题自足句。定义详见文献[3]。

[2] “北京语言大学语言信息处理研究所广义话题结构标注语料（2014 年 5 月 28 日版）”共有 37 万余字，含 3 万多个标点句，包括小说、百科全书、政府工作报告等多种语体的语料，详见文献[9]。语料免费公开使用，地址为：http://pan.baidu.com/s/1i3qpibb。

的区别意义。有些动词在语料库中出现频率很高，但没有引出过新支话题，如“去”“没有”“工作”等。语料中，引出新支话题数量最多的动词是“有”（84 例）、“是”（46 例）、“为”（26 例），其他都不超过 20 例。

3.2. 接续特征

接续特征的计算方法是：以二元语言模型为基础，分别计算上句动词宾语与本句首词接续概率，和主语与本句首词分别的接续概率，再以两者概率之差为判别新支样本的候选特征。

汉语缺少主谓一致等形态变化（性、数、格等），无法通过语言中的标记来判断话题接续，故我们通过接续概率来预测。如果样本中本句的话题是上句主语/宾语，那么对于汉语来说，上句主语/宾语与本句通常可以不经删改直接连成句法通顺、语义合理的句子。既然如此，这种接续情况在大规模语料下应该会出现在一个完整的句子中，接续概率相对高。反之，若上句主语/宾语与本句首词接不起来，那么这种接续概率在大规模语料库中很可能较低。例如：

例 2：他前面放个滚圆的麻袋，
‖ 里面想是米

该例是新支样本的例子，上句主语是“他”，宾语是“滚圆的麻袋”，统计时以“麻袋”为宾语核心词。本句首词是“里面”。从接续概率上看，“麻袋里面”比“他里面”概率高，故按接续概率计算，倾向于将该例判断为新支样本。

由于句法的多样性和复杂性，目前汉语自动句法分析的准确率不高，故在计算时，对上句宾语，我们通过人工标记出每个样本上句动词宾语的核心成分，再在语言模型中查询该成分和本句首词的接续概率；对上句主语，我们既不做句法分析，也不做人工标注主语核心词，而是直接分词处理，计算上句动词前每个词与本句首词的接续概率，取其最大值作为该样本上句主语和本句的接续概率。

3.3. 信息量特征

文献[11]指出，信息量对新支句的形成有重要作用。若上句宾语的信息量越小，则越有必要对宾语所指事物进一步说明，故宾语成为新支话题的可能性越大。反之，若上句宾语信息量越大，则对宾语所指事物再加以说明的必要性越小，宾语成为新支话题的可能性越小。此时，本句倾向于说明主语的某些结果，上句主语成为本句话题的可能性大。

由于信息量不容易计算，在实验中，我们用了两个特征进行粗糙的模拟，即宾语词数和字数比所有样本宾语平均词数和字数多还是少。

3.4. 句法特征

这里的句法特征是指能够从字面识别的特征，这些特征是人的认知和语义在字面上的体现。这类特征可以被计算机直接识别，包括：

1) 标点符号特征：若上句标点为句号、叹号、问号、分号，本句一般不成为新支句。

上句句末为句号、分号、叹号、问号等有较大停顿的标点符号时，表示上句意义相对完整，故本句难以就上句的宾语作进一步说明。

例 3：a）颚针鱼科下咽骨被有细小尖齿；
鼻骨大，
b）颚针鱼科两颌具细小尖齿，
‖ 呈带状排列，

上例都是百科全书中的原文，例中 a 和 b 的上句意思相当，宾语相同。a 的上句句末用分号，提示读者，后文不会再对“细小尖齿”做进一步说明。b 的上句句末用逗号，提示后文可能继续说明“细小尖齿”。

2) 本句句首是后连词时，一般不成为新支句。

例 4：王脚擦汗时看到儿子王肝和女儿王胆，
便大声喝斥

该例的上句主语“王脚”和宾语“儿子王肝和女儿王胆”，在语义上都能与本句“大声

喝斥”搭配。位于标点句句首的连词“就、便”等表示上文动作的顺承。因此，本句话题倾向于“王脚”。

3) 上句特征动词后有趋向动词时，其动词宾语倾向于成为新支话题。

趋向动词往往表示句中主体（人或事物）的位置移动[12]，若移动的主体是宾语时，该主体往往因位置移动而从隐蔽处显现，有进一步说明的需要，容易成为新支话题。

例 5：阿刘手向口袋里半天掏出来一只发钗，

‖ 就是那天鲍小姐掷掉的。

例中“发钗”是因位移引介出来的新事物，成为了新支话题。

4) 本句是关系句而上句不是关系句时，本句倾向成为新支句。

关系句的判断以一系列关键词作为代表，包括表判断的“是、属于、当作、称为”等，表相似比喻关系的“像、好像、比如”等， 以及表比较的“比”等。如：

例 6：老大这个孩子后来看中苏鸿业的女儿，

‖ 也是有钱有势的人家。

该例上句不是关系句，本句是表归类的关系句，是对上句宾语“苏鸿业的女儿”描写，成为新支句。

5) 本句是有字句而上句不是有字句时，本句倾向成为新支句。

例 7：车拉到法租界边上，

‖ 有一个法国巡捕领了两个安南巡捕在搜检行人，

“有”字的一大功能是表存在。该例动词宾语“法租界边上”是方位短语，与本句构成存现句，故本句成为新支句。

例 8：沙发旁一个小书架猜来都是张小姐的读物。

‖ 有原文小字白文《莎士比亚全集》、《新旧约全书》、《家庭布置学》、翻版的《居里夫人传》、《照相自修法》、《我国与我民》等不朽大著以及电影小说十几种

该例的本句表列举，是对动词宾语的外延的进一步扩充。

3.5. 语义特征

语义特征是判断是否新支样本的最主要因素。实际上，上文的接续特征本身就是语义特征的一种，它模拟了人的认知中某两个词语之间的紧密程度。这里的语义特征专指需要人工语义标注的特征。

1) 语义泛化

实验中，我们对上句主/宾语核心词、本句主语核心词以及本句谓语核心词进行了人工语义泛化标注。主/宾语核心词的语义泛化类型有：人，人的部件，人的部位，人的属性，人的反应性部位（如“心里”），事物，事物部位，事物部件，事物属性，书信，书信部件，书信属性，信息，指示词，抽象物。

对本句谓语核心词的语义泛化类型有：一般行为，反应性行为，反应性形容词，一般形容词，状态动词，具有“是”“有”“说”“看”“听”意义的动词，关系动词，一般名词。

2) 基于语义泛化的平行结构

若经过语义泛化后，上句中存在某一个后段与本句结构相似，则标记该样本为平行结构。此时，本句共享上句平行结构前的部分作为话题。为避免句法分析，平行结构均人工标注。

例 9：阿古柏本为浩罕的军官，

初为浩罕国王呼达雅尔汗的”穆合热本”，

该例上句和本句构成“时间副词+担任+隶属者+职务”的平行结构，本句共享“阿古柏”为话题。

例 10：自蓟城向南可直下中原，

向西北径上蒙古高原，

该例上句和本句构成“自+处所 A+向+方向+到达+处所 B”的平行结构，本句共享“自

蓟城”为话题。

3.6. 其他特征

除了以上列举的特征外，常识和专业知识对新支样本的判断也有影响。

例 11：他们路上碰见两个溃兵，

‖ 抢去方老先生的钱袋，

从语义上看，“他们抢去了钱袋”也是通顺的。但是常识上“溃兵”容易让人与“抢”的施动者联系起来，故人倾向将本句理解成新支句。

例 12：鲹尾鳍分叉深，

下叶比上叶略长；

若不具备专业知识，就不知道“下叶”是“尾鳍”的组成部件，还是“鲹”的身上与“尾鳍”同等地位的部件。故本例的判断使用了鱼类部件的专业知识。

但是，由于常识和专业知识类特征过于复杂，难以提取，本文实验暂没采用。

在以上特征中，动词特征、信息量特征、接续特征、句法特征都是可以通过字面统计或推导出来的，实验中统称为字面特征，而区别于需要人工标注的语义特征。

4. 实验方法

4.1. 模型的选择

通过上节的分析可见，判别新支样本的特征是分别从句法、语义、语用甚至常识中提取出来的，这些特征的粒度差异大，特征间的同质性不高，难以预测在自然语言中的概率分布，故我们采用最大熵模型进行机器学习模型。因为最大熵模型的特征选择较为灵活[13]，且特征之间不需要的独立性假设或者其他内在约束，能够较好地把人的知识以特征的形式融合到统计模型中，最大限度将人的知识与统计方法相结合。

4.2. 模型的调整

由于最大熵模型是以整体准确率来评价结果的好坏的，而新支句判别问题是一个非均衡的分类问题。总体样本中，新支样本 641 例，非新支样本 1508 例，比例约为 1:2.35。这种情况下，即使把全部样本 2149 例全部判为非新支句，整体准确率也能达到 70.2%。但这并非实验所要达到的目标。我们更关注新支样本的准确率和召回率。故在实验中，我们调整新支样本的权重。方法是：在构造训练集时，将新支样本复制若干份，使得新支样本和非新支样本比例约为 1:1 左右，而测试集则保持原来的比例不变。

4.3. 测试方法

由于总体样本较少，为了更充分利用有限的样本，我们采取“留一交叉验证”的方法进行测试。具体方法如下：将非新支样本集和未经复制的新支样本集合起来作为“**原始库**”；将非新支样本集和复制了若干份的新支样本集合起来作为“**调整库**”。每次实验，在原始库中取一个样本作为唯一测试对象，调整库中临时除去这个测试对象作为训练集。如此，对原始库中的每个样本都测试一次，最后对原始库所有样本的测试结果进行统计。

5. 实验过程和结果

5.1. 实验一（Baseline）：基于字面特征的全语料新支判别

原语料中，新支样本 641 例，非新支样本 1508 例。训练时，经权重调整，新支样本调整为原来的 3 倍，即 1923 例；测试时，按原语料逐一进行留一交叉验证。由于时间和精力有限，全语料的判别实验仅采用字面特征，包括：动词特征、接续特征、句法特征和信息量特征，没有引入语义特征。实验结果如表 1 所示。

表 1 基于字面特征的全语料新支判别结果

新支总数	641	
新支错误	241	37.60%
新支正确	400	62.40%
非新支总数	1508	

非新支错误	436	28.91%
非新支正确	1072	71.09%
样本总数	2149	
总错误率	677	31.50%
总正确率	1472	68.50%
新支准确率		47.85%
新支召回率		62.40%
新支 F 值		54.16%

注：表中“新支正确”指的是新支样本被判为新支；“新支错误”指的是新支样本被判为非新支；“非新支正确”指的是非新支样本被判为非新支；“非新支错误”指的是非新支样本被判为新支。下同。

这一结果是新支样本自动判别实验的 Baseline。新支样本判断的准确率为 47.85%，召回率为 62.40%，非新支样本的判断正确率要高于新支样本近 10 个百分点。为衡量各个特征的贡献度，我们计算了每个特征的信息增益（表 2）以及各特征权重值 λ（表 3）。

表 2 全语料中各特征的信息增益（前 5）

增益排序	特征说明	特征名	信息增益($\times 10^{-4}$)
1	上句核心谓语动词	yyv_fea	2343
2	标点符号	finish_fea	933
3	本句是关系句	jf-gxj	288
4	主/宾语与后句首词接续值	lj_fea	111
5	上句宾语的字数	xxlz_fea	38
……	……	……	……

表 2 是按照调整库来计算特征的信息增益，按照信息增益的值由大到小列出了特征信息增益前 5 的特征。其中，区分度最明显的是上句核心谓语动词的特征，远高于其他特征。其次是标点符号。

表 3 全语料中非动词特征的权重值 λ

权重排名	特征值	特征权重值
1~87	动词特征	……
88	句号类标点	lambda(0,FinishedSent)=0 lambda(1,FinishedSent)=1.46038
91	本句句首是“就/便”	lambda(0,jf-hlc)=0 lambda(1,jf-hlc)=1.40833
113	本句是关系句	lambda(0,jf-gxj)=1.18948 lambda(1,jf-gxj)=0
164	上句核心谓语动词后是趋向动词	lambda(0,jf-qx)=0.638714 lambda(1,jf-qx)=0.0233809
182	逗号冒号类标点	lambda(0,UnfinishedSent)=0.455184 lambda(1,UnfinishedSent)=0
……	……	……

注：（1）表中的 lambda 函数是最大熵实现程序中的特征权重的表现形式。lambda 函数的参数中，第一个参数值 0 表示新支，1 表示非新支，第二个参数是特征值，如 FinishedSent 表示句号类标点。lambda 函数表明经过模型训练后，该特征倾向于对新支样本还是非新支样本有贡献，等号后的数字表示其权重。下同。（2）动词特征权重表中没有列出，如排名在 1~87，89~90 等的特征，均为动词特征。

表 3 列出了排在前 5 名的非动词特征值。本实验中，含动词在内全部特征值有 248 个，而排在前 88 的都是动词特征（即各动词词形，表中没有列出），可见动词特征是影响新支判别最重要的因素。除动词特征外，上句句末的标点符号为句号、叹号、问号等标点符号作为特征值的权重最大，从 lambda 函数看，模型认为遇到这类特征倾向于判断为非新支样本。其次是本句句首为“就”或“便”这种后连词，模型倾向于判断为非新支样本。

结合以上两个表可见，动词特征对新支样本的判别效果是最显著的。故，为了进一步考察不同动词对于其他各种特征及特征值敏感程度的差异，我们选择了两类有代表性的典型动词进行实验。一类是动词“有”；另一类是“看听”类动词，包括“看”、“听”、“见”、“瞧”、“看见”、“听见”、“瞧见”、“看看”等。我们把含有这些典型动词的新支样本和非新支样本挑出来，单独组成该类实验动词的语料库，进一步做语义泛化标注和实验。

5.2. 实验二：动词“有”类样本的单独实验

挑选“有”作为典型动词进行实验，有以下几个原因：从统计上看，动词“有”新支样本数量最多，且总体样本数量也最多，有较好的统计意义。从语义上看，“有”的义项中出现最多的是领有和存在，它们的语用意义很多情况下是引出上文中未出现过的新事物，很可能接下来要介绍这个新事物，因此“有”的宾语成为新支话题的可能性大。在语料中，含有“有”的新支样本共 84 例，非新支样本 446 例。训练时，调整库中，新支样本调整为原来的 6 倍，即 504 例；测试时，按原语料逐一进行留一交叉验证。我们首先做基于字面特征的实验，然后加入语义特征再次实验。

（1）基于字面特征

对“有”的实验，按照接续特征、句法特征、信息量特征等字面特征进行最大熵的训练，结果如表 4 右栏。

表 4 “有”类样本在全语料实验中和“有”的单独实验中新支判别结果比较

	全语料实验		“有”的单独实验	
新支总数	84		84	
新支错误	59	70.24%	8	9.52%
新支正确	25	29.76%	76	90.48%
非新支总数	446		446	
非新支错误	36	8.07%	179	40.13%
非新支正确	410	91.93%	267	59.87%
所有样本总数	530		530	
总错误	95	17.92%	187	35.28%
总正确	435	82.08%	343	64.72%
新支准确率		40.98%		29.80%
新支召回率		29.76%		90.48%
新支 F 值		34.48%		44.84%

全语料中“有”样本判别结果和“有”单独实验相比，二者选取的特征是相同的。全语料实验中，新支样本的正确率只有 29.76%，模型把大部分“有”类样本判定为非新支样本，包括 410 个非新支样本和 59 个新支样本，共 469 个，占全体 530 个样本的 88.5%。而“有”类样本单独实验中，模型把大部分新支样本都判断正确了，新支样本正确率 90.48%。但也把 179 个非新支样本判为新支样本。我们考察“有”类样本单独实验的特征权重值：

表 5 “有”类样本单独实验中字面特征统计的特征权重值 λ

权重排名	特征说明	特征权重值
1	本句是关系句	lambda(0,jf-gxj)=1.97322 lambda(1,jf-gxj)=0
2	句号类标点	lambda(0,FinishedSent)=0 lambda(1,FinishedSent)=1.90353

权重排名	特征说明	特征权重值
3	本句是有字句	lambda(0,jf-you)=0 lambda(1,jf-you)=1.34257
4	逗号冒号类标点	lambda(0,UnfinishedSent)=0.83122 lambda(1,UnfinishedSent)=0
5	本句句首是“就/便”	lambda(0,jf-hlc)=0.608869 lambda(1,jf-hlc)=0
……	……	……

对比表 3 和表 5，两个实验使用了相同的字面特征，但是权重值排序不一样，有些具体的特征倾向性也不一样。因为在全语料中，“有”的样本的判断正确率受到其他动词的干扰。

（2）基于字面特征+语义泛化

“有”类样本的语义特征标注包括平行结构和宾语语义泛化。

平行结构定义如 3.5 节，具有平行结构的样本，本句倾向于成为非新支句。

“有”的宾语语义泛化可以分为两类，一类指具体事物，一类指抽象事物。具体事物较容易作为新支话题，而抽象事物作新支话题通常比较困难。如：

例 13：１９３０年时仅于芜湖**有**纺织厂，

‖ 规模均很小

上例“纺织厂”是一个具体的事物，被“有”引出后，从认知上，有需要介绍其更多情况，如规模、产量、产品等属性。

例 14：罗素在国际上享有声望。

曾任国际天文学联合会恒星光谱组和恒星结构组主席。

上例“声望”是一种抽象的概念，内涵比较单一且明确，被“有”引出后，不需要对其属性进一步说明。

这两种特征引入模型后，含语义特征在内的各特征的信息增益情况如表 6。在“有“的语料中，所有平行结构的样本都为非新支样本，而上句宾语为抽象名词的样本也大多数是非新支样本，故这两种特征的信息增益都较大。

表 6 “有”类样本单独实验增加语义特征后信息增益（前 5）

增益排序	特性说明	特征名	信息增益($\times 10^{-4}$)
1.	标点符号	finish_fea	1955
2.	平行结构	px	995
3.	上句宾语的语义泛化	obj_fea	763
4.	本句是关系句	jf-gxj	486
5.	本句是有字句	jf-you	418
……	……	……	……

实验结果如下。

表 7 “有”类样本单独实验增加语义特征后的新支判别结果

	不添加语义特征		添加语义特征	
新支总数	84		84	
新支错误	8	9.52%	8	9.52%
新支正确	76	90.48%	76	90.48%
非新支总数	446		446	
非新支错误	179	40.13%	104	23.32%
非新支正确	267	59.87%	342	76.68%
所有样本总数	530		530	
总错误	187	35.28%	112	21.13%

	不添加语义特征	添加语义特征
总正确	343 64.72%	418 78.87%
新支准确率	29.80%	42.22%
新支召回率	90.48%	90.48%
新支 F 值	44.84%	57.58%

表 7 显示，添加语义特征后，召回率不变，而准确率提高了 60%。可见平行结构和宾语语义泛化作用明显。但仔细考察新支判别错误的例子发现，虽然都是 8 个错误，但分别有 4 个样本在不添加语义特征时判断正确的，添加语义后判断错了，还有 4 个样本是不添加语义时判断错误而添加语义特征后判断正确。如：

例 15 不添加语义特征时判断正确，添加语义特征后反而判断错误

解决台湾问题可以有两种方式，

‖ 一种是非和平的方式，

例 15 的“方式”是抽象名词，由于语义特征的重要影响，根据特征的信息增益和模型的权重，倾向于判为非新支样本。但仔细分析例 15 错判的原因发现，虽然“方式”是抽象名词，但其前面有数量短语“两种”。通常数量名短语作为句末的宾语时，有进一步解释的需求。而这个特征实验中之前没有发现。可见，特征选取还有很大的研究空间。但是，特征越多、越细，样本数据就越稀疏，越可能发生过度拟合，这是另一个令人纠结的困难。

表 8 “有”样本单独实验增加语义特征后各特征权重值 λ（前 5）

排名	特征说明	特征权重值
1.	平行结构	lambda(0,px)=0 lambda(1,px)=2.45251
2.	句号类标点	lambda(0,FinishedSent)=0 lambda(1,FinishedSent)=1.78022
3.	本句是关系句	lambda(0,jf-gxj)=1.71218 lambda(1,jf-gxj)=0
4.	上句宾语是抽象名词	lambda(0,obj_abs)=0 lambda(1,obj_abs)=1.69554
5.	本句是有字句	lambda(0,jf-you)=0 lambda(1,jf-you)=0.905809
……	……	……

5.3. 实验三：“看听”类动词样本的单独实验

“看听”类动词语义上通过感官的认知引入一个对象，这个对象通常是较为具体的对象，如一个人、一个物体，一条消息等，故有深入介绍其特性或内容的需要。实验所用“看听”类动词包括：看，看见，看到，看得（“他看得几页”），细看，偷看，瞧，瞧见，瞧着，见，听，听见，听到，听清，听说，碰到，碰见。在语料中，含“看听“的新支样本有 62 例，非新支有 101 例。训练时，把新支样本调整为原来的 2 倍，即 124 例，非新支 101 例保持不变。测试时，仍使用留一交叉验证。

（1）基于字面特征

对“看听”的样本，首先按照接续特征、句法特征、信息量特征等字面特征进行最大熵的训练，不包括语义泛化的特征，结果如下：

表 9 “看听”类样本基于字面特征的新支判别结果

新支总数	62	
新支错误	12	19.35%
新支正确	50	80.65%
非新支总数	101	

非新支错误	33	32.67%
非新支正确	68	67.33%
样本总数	163	
总错误率	45	27.61%
总正确率	118	72.39%
新支准确率		60.24%
新支召回率		80.65%
新支 F 值		68.97%

对“看听”类样本而言，仅基于字面特征的效果已经达到 60.24%的准确率和 80.65%的召回率。可见，实验选用的特征，比较适合判别“看听”类动词引起新支话题。

表 10 “看听”类动词样本单独实验中各字面特征权重值 λ（前 5）

权重排名	特征说明	特征权重值
3	本句首词“就/便”	lambda(0,jf-hlc)=0 lambda(1,jf-hlc)=2.0945
5	本句是有字句	lambda(0,jf-you)=1.78423 lambda(1,jf-you)=0
8	本句是关系句	lambda(0,jf-gxj)=1.15326 lambda(1,jf-gxj)=0
11	上句宾语的字数低于平均值	lambda(0,xxlz-l)=0 lambda(1,xxlz-l)=0.895405
13	上句宾语的字数高于平均值	lambda(0,xxlz-h)=0.7625 lambda(1,xxlz-h)=0
……	……	……

表 10 列出了除动词特征外，字面特征的权重值排前 5 的特征。可以看出，排在前列的还是以动词特征居多，但前几个实验中区别显著的标点符号类特征并没有排在前列。

（2）基于字面特征+语义泛化

仅有上述一些特征，显然不足以描述新支话题的形成原因，进一步，我们针对动词的主语、宾语和本句核心动词以及本句句首副词或主语，进行语义泛化。泛化内容如 3.5 节。

表 11 “看听”类样本统计+语义的新支判别结果

新支总数	62	
新支错误	8	12.90%
新支正确	54	87.10%
非新支总数	101	
非新支错误	17	16.83%
非新支正确	84	83.17%
总数	163	
总错误率	25	15.34%
总正确率	138	84.66%
新支准确率		76.06%
新支召回率		87.10%
新支 F 值		81.20%

加入对上句主/宾语、本句主/谓语的人工语义泛化的标注后，效果有了明显的提升，召回率达到 87.10%，准确率也达到 76.06%。

表 12 “看听”类样本特征信息增益（前 5）

增益排序	特征说明	特征名	信息增益($\times 10^{-4}$)

增益排序	特征说明	特征名	信息增益($\times 10^{-4}$)
1.	本句核心谓语语义泛化	bkv_fea	4122
2.	上句核心谓语动词	yyv_fea	2349
3.	本句句首是反应性副词	bkd_fea	1729
4.	上句宾语的语义泛化	obj_fea	1698
5.	本句主语语义泛化	bkn_fea	1633
……	……	……	……
15.	标点符号	finish_fea	11

表 12 显示了包括语义泛化后各特征的信息增益。本句核心谓语和上句核心谓语相关的特征信息增益都较大，在“有”类实验中作用显著的标点符号增益最小。进一步考察各特征的权重值：

表 13 “看听”类样本各特征权重值 λ

权重排名	特征说明	特征权重值
1	本句为主语是指示代词	lambda(0,bk_prop)=0 lambda(1,bk_prop)=2.51201
2	本句谓语是听	lambda(0,bk_hear)=0 lambda(1,bk_hear)=2.3121
3	本句主语是身体部位	lambda(0,bk_hum-pos)=0 lambda(1,bk_hum-pos)=2.10557
4	本句主语是人的属性	lambda(0,bk_hum-attr)=0 lambda(1,bk_hum-attr)=1.9633
5	本句主语是人的反应性部位	lambda(0,bk_reacp)=0 lambda(1,bk_reacp)=1.88814
……	……	……

表 13 显示，权重靠前的特征都是语义泛化特征，其作用还大于动词特征。而且，本句相关的语义特征比上句有关的语义特征作用更明显。

有意思的是，比较“看听”类样本“基于字面特征”和“基于字面特征+语义泛化”两组实验中的新支错误的数据时发现，原来“基于字面特征”的 12 个新支样本判断错误，经过语义泛化，“基于字面特征+语义泛化”中有 9 个判断正确了，但却有 5 个原来判断正确的新支样本，语义泛化后反而判断错了。这 5 个例子如下：

例 16

（1）有一天魏队长看着那老汉扬着鞭子，
‖ 一副怡然自得的样子，

（2）天赐细看自己，
‖ 确是身量高了

（3）他碰见李妈，
‖ 正要说话，

（4）直到傍晚鸿渐碰见她，
‖ 说正要来问赵叔叔的事。

（5）我细看他相貌，
‖ 也还是乱蓬蓬的须发；

例中列出了各个例子的特征和具体例子，（1）~（4）主宾语都是人，（5）的主语是人，宾语是人的属性，在语义上，上句的主语和宾语基本没有区别，模型没能判断孰优孰劣。

再看非新支判断错误的例子，有些是不应该判断错的，如：

例 17：他看得几页，

眼前金光一闪，

系统把该例判断成了新支样本，而例中，上句主语是人，宾语是书信类，本句句首“眼前”是人体部位，而“书信”是无法和人体部位相连接的。但由于把语义泛化作为特征时，并没有考虑上句主语、宾语语义和本句句首主语或者谓语的接续关系，所以这种不合理的接续未被发现。而由于实验语料太少，学习这种接续关系，将面临数据严重稀疏的问题。因此，下面我们使用规则的办法，把这种不可能相接的关系作为规则引入判断体系中。

（3）基于字面特征+语义泛化+规则

计算机只能根据概率给出答案，但无法断言某种答案不可能存在，只能指定小概率的范围。因此通过人为给出规则判定，可以帮助计算机提高性能。具体方法是：把上句主语、宾语的语义类型和本句句首、本句主语、本句核心动词的语义类型一一比对，根据人的认知：将不可能匹配的语义二元组建立为否定规则，实验中遇到满足否定规则的情况，直接确定相反的情况为判断结果。对于不满足否定规则的情况不做判断。语义接续否定规则举例如下：

表 14 语义接续否定规则（举例）

上句主/宾语语义类型|本句语义类型	上句本句符号表示
抽象事物|书信	abs|book
抽象事物|书信属性	abs|book-attr
……	……
动物|人	ani|hum
动物|人的属性	ani|hum-attr
动物|人的部件	ani|hum-part
动物|事物部件	ani|thing-pos
……	……
书信|听	book|hear
书信|反应性形容词	book|reaca
书信|反应性副词	book|reacd
……	……

例如，例 8.8 的样本，上句宾语“几页”泛化成“书信”，本句主语（亦即首词）“眼前”泛化为“人体部位”，“书信”和“人体部位”满足否定规则，直接判为不可能发生新支，于是只能判为非新支。

加入规则后，我们的实验方案修改为：先通过规则，把能够判定的先判定，不能够判定的交给最大熵模型处理。实验结果如表 15 所示，并跟没有添加规则的结果（表 11）相比较：

表 15 “看听”类样本单独实验中统计+语义+规则的新支判别结果

	统计+语义		统计+语义+规则	
新支总数	62		62	
新支错误	8	12.90%	8	12.90%
新支正确	54	87.10%	54	87.10%
非新支总数	101		101	
非新支错误	17	16.83%	15	14.85%
非新支正确	84	83.17%	86	85.15%
所有样本总数	163		163	
总错误	25	15.34%	23	14.11%
总正确	138	84.66%	140	85.89%
新支准确率		76.06%		78.26%
新支召回率		87.10%		87.10%
新支 F 值		81.20%		82.44%

可以看出，添加语义规则后，对新支判断没有影响，对非新支的错误数从 17 例下降到 15 例，有 2 例非新支原来判断错误的，现在正确了，如下：

例 18

（1）他只看得几页，

不由得吓了一跳，

（2）我们看过这面“治岗红旗”，

心里都非常感奋，

上例中，（1）句的宾语“几页”是书信类，本句首词“不由得”是反应性副词，不可能相接。（2）句的宾语“这面‘治岗红旗’”是事物，本句首词“心里”是人反应部件，不能相接。这两个例子是规则判断的结果。

5.4. 实验四：含“看、听、有”语料的新支判别

在以上实验的基础上，我们尝试把“看听”、“有”两类动词样本综合起来，考察它们的表现情况。由于“有”和“看听”使用的语义泛化方法不一样，故本实验仅使用基于字面特征的方法进行训练和测试，不加入语义泛化的特征。“看听”、“有”共有新支样本 146 例，非新支 547 例，训练时，把新支语料调整为原来的 4 倍，即 584 例，非新支语料 547 例不变。测试结果如下：

表 16 “看听有”类样本基于字面特征的新支判别结果

新支总数	146	
新支错误	40	27.40%
新支正确	106	72.60%
非新支总数	547	
非新支错误	205	37.48%
非新支正确	342	62.52%
样本总数	693	
总错误率	245	35.35%
总正确率	448	64.65%
新支准确率		34.08%
新支召回率		72.60%
新支 F 值		46.39%

从表现测试结果看（表 16），准确率在“有”和“看听”类样本实验之间，但是，召回率却比“有”和“看听”类样本实验都要低。可见，两类动词由于表现不一样，需要的特征和权重不一样，把他们混到一起会出现两类特征出现相互制约的情况。这正是全语料字面特征实验（Baseline）中，结果不太好的一大原因。

6. 讨论

通过新支样本和非新支样本的判别实验，我们尝试将统计方法和认知规则及人的语义知识结合起来进行判定。总的来说，自动判别是比较复杂的。

表 17 各实验结果比较

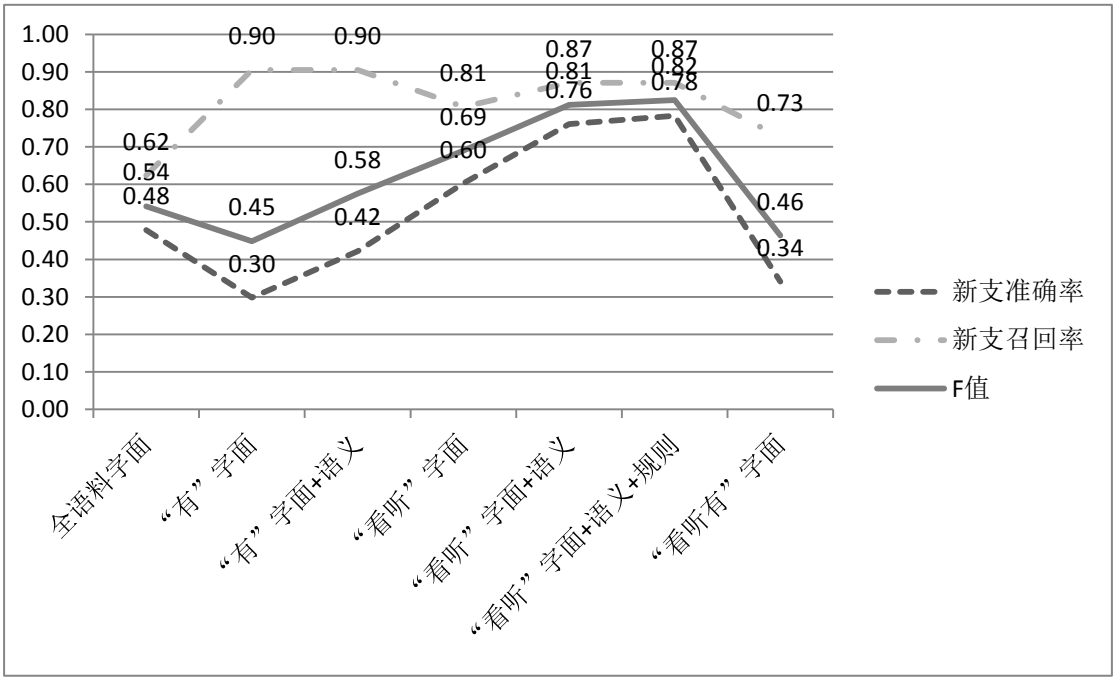

表 17 比较了各实验的新支准确率、召回率和 F 值，由实验可知：

（1）动词特征起重要作用。一方面，动词特征在训练后，权重皆比较高；另一方面，同样的特征，对不同动词专门训练，所得到的其他特征的权重也不一样。动词特征的重要作用从另一个方面说明，对不同的动词应该使用不同的特征进行判定。反推人的认知，我们提出一种可能性，即人在判断后标点句的话题是上句的主语还是宾语时，也是根据动词的语义、语用等因素，调取不同的模板进行判断，而不是通过同一套特征及其权重来判定。

（2）语义泛化对新支判断的影响重大。凡经过泛化，都能使得效果有较大提高。而且语义泛化特征的权重都排在较前的位置，证明语义泛化标注具有较好的一致性。语义泛化需要人的知识的介入和标注，再辅以统计学习方法才能获得较为良好的效果。

（3）规则能够帮助提高判断的准确率。如果仅靠规则，由于变化因素多，相互关系复杂，写规则时难以面面俱到，准确无误地描述。但是统计的一大缺点是无法对否定进行断言，只能按照一个小概率的范围来估计和拒绝。如果能总结出不可能的规则，则能够帮助计算机提高效率和准确率，减少对不可能的事情的错误估计。

诚然，本实验还能有许多能够改进的地方。

（1）有几个特征的获取依赖于句法分析，如果提高句法分析的准确性，可以减少人工标注，获得更多的训练数据。如主宾语核心成分的提取，以及平行结构的识别，它们都在特征中起到重要作用。

（2）有些特征的计算方法不科学，比如宾语信息量的特征，作用甚微，甚至有时起到反作用，应重新设计计算方法。

（3）统计方法上，不一定只选择最大熵模型，可以结合多种方法进行尝试，本文由于时间关系，没有开展更多的实验。

从本文的实验及其分析，我们认为得出以下几点结论：

（1）统计方法和规则方法不能偏废。随着大数据的兴起，深度学习的出现，学界对统计方法有了新一轮的期待。越来越多的声音认为统计方法将能取代规则或者自动发现规则。规则的作用在于断言，能够把不可能的情况排除在外。我们的实验证明，规则确实能够提高系统的性能，系统的设计应留有接口，介入规则。

（2）注意精细知识的使用。本实验的一条重要的结论是，不同动词适用不同的特征。这就要求对特征的描述非常准确和精细，对每类动词，应根据其语义、语用、认知等构造一套语义特征模板。这不能缺少人的参与。

（3）统计模型和人的作用并重。统计中的特征选取，包括字面特征和精细的语义泛化，以及规则的确定等，这一系列的过程都不能离不开人的参与。人在认知时，依赖于许许多多的知识模板，这些精细的模板必须由人来提供一定的知识支持，再辅以统计模型，才能取得更好的效果。因此，自然语言处理中，不仅不能忽视人的因素，还需要有大量深入的人的智力投入，深入到语言事实语言现象中，发掘和思考认知原理。

7. 结语

本文尝试让计算机自动判别标点句所缺的话题。考虑到任务的复杂性，最后限制在上句主语和宾语的判别上，即仅区分新支样本和非新支样本。主要工作内容是实验语料的获取、统计模型的确定、特征的选取、实验的组织。实验组织中涉及到不同对象语料、不同特征类的多种组合以及规则的加入。实验结果是：仅用字面特征的全语料的最大熵模型计算中，新支句判断的 F 值为 54%，对于“看听”类动词的样本单独实验，加入比较丰富的语义特征并使用否定型的规则后，新支句判断的 F 值达到 82%。实验说明，即使在有限范围内的自动识别，工作难度也较大，且严重依赖于人的语言知识。

本文的实验只是进行初步的探索，而且由于时间关系，实验过程使用的特征和模型参数还比较粗糙，本实验的目的并不在于令标点句缺失话题的自动判别达到实用化，实验结果并非十分理想，但实验证明，基于把统计模型和认知方法相结合是可行的，其结果的正误是基本可解释的。

参考文献

[1] 宋柔. 汉语叙述文中的小句前部省略现象初析[J]. 中文信息学报, 1992, 6(3):62-68.

[2] 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, 2008:26-44.

[3] 宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文, 2013(6):483-494.

[4] 卢达威，宋柔，尚英. 从广义话题结构考察汉语篇章话题的认知复杂度[J]. 中文信息学报, 2014(5):112-124.

[5] 卢达威. 语料库实证的汉语篇章广义话题结构认知和计算研究[D]. 北京语言大学博士论文, 2015.

[6] 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] 蒋玉茹, 宋柔. 基于细粒度特征的话题句识别方法[J]. 计算机应用, 2014, 34(5):1345-1349.

[8] 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.

[9] 尚英. 汉语篇章广义话题结构理论的实证性研究[M]. 北京语言大学博士论文, 2014.

[10] 季翠, 卢达威, 宋柔. 动词引出新支话题的语用功能研究[J]. 中文信息学报, 2014(3):22-27.

[11] 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[M]. 中国社会科学出版社, 2013.

[12] 张斌. 现代汉语描写语法[M] 北京：商务印书馆, 2010.

[13] Berger, Adam L; Pietra, Vincent J. Della; Pietra, Stephen A. Della. A Maximum Entropy Approach To Natural Language Processing [J]. Computational Linguistics, 1996, 22(1):39--71.

作者简介：卢达威（1983——），男，博士，主要研究领域为语言信息处理。Email:wedalu@163.com；宋柔（1946——），男，教授，博士生导师，主要研究领域为语言信息处理。 Email:songrou@126.com。


（卢达威）


（宋柔）