汉语“比”字句要素的常规序列模式探索 *

朴敏浚　李强　袁毓林

（北京大学 中文系/中国语言学研究中心/计算语言学教育部重点实验室，北京 100871）

摘要：表达“差比”义的“比”字句，是比较句的重要句型，也是比较句关键要素抽取问题中不可回避的主要句型。该句型的关键要素（SUB、BI、OBJ、ITM、DIM、RES、EXT）在语义上互相交织，在表层句法上可以实现为多种多样的序列模式。面向中文“比”字句关键要素抽取问题，本文对于表示“差比”义的 460 多个“比”字句文本进行了七种关键要素的标注。在此基础上利用 Apriori 和 PrefixSpan 算法找出这些要素的关系规则及其序列模式，并归纳出六种“比”字句关键要素的分布规律。此外，本文还进一步说明产生这六种模式规则的动因，对于“比”字句特征选取问题提供重要的语言学的启发以及理论依据。

关键词：“比”字句；关键要素；关系规则；序列模式；分布规律

A study on the Sequential Patterns of Semantic Constituents of the Bi-Comparative Structure

Abstract: The Bi-structure, which highlights a contrasting characteristic between two elements, is the key comparative sentence structure in Chinese. Therefore, it has been the main target of keyword mining of Chinese comparative sentence. This structure consists of 7 types semantic items (SUB, BI, OBJ, ITM, DIM, RES, EXT), of which various sequential patterns may occur. To provide meaningful information for the keyword extraction task of this comparative structure, this study first begins with the tagging of the 7 semantic items on about 460 sentences. Second, associative rules and sequential patterns are extracted using the Apriori and PrefixSpan algorithms, from which 6 rules of the item distribution are established. Finally, this paper illustrates the rationale behind the construct of these 6 rules, providing a better understanding of the particular characteristics and useful insight for feature selection task of the Bi-comparative structure in Chinese.

Keywords: Bi-structure; keyword extraction; sequential pattern mining; distribution rule

1 引言

随着互联网的普及，社交网络（SNS）的影响力也随之日益增强，成为人们的信息交流平台，这自然而然地引起了企业对网络大数据（big data）进行分析与探索的动机和兴趣。其中，与商品销售有着直接关系的网络商品评价，尤其是顾客评价中的比较句成了信息抽取（IE）领域的研究热点。比如，黄小江（2008）提出以类序列规则（CSR）为特征的汉语比较句识别模型。张晨等（2013）利用了类序列规则（CSR）、语义角色信息、统计词等多种特征，通过模板匹配和机器学习相结合的方法提高比较句识别表现。此外，中文倾向性分析评测（COAE）连续两年（2012、2013）把比较句的识别与要素抽取列为评测任务。大部分参赛单位在特征词序列模式的基础上利用模式匹配的方式进行了关键要素提取（侯名牛 等 2013；李岩 等 2013）。虽然这多方面的努力完善了比较句的识别表现，但现有的比较关键要素提取模型的表现还停留在初步水平 [1]。这是因为关键要素的提取是覆盖分词、短语识别、未登录词识别、句法分析等多方面的技术（谭松波 2013：32）。然而，我们认为目前最关键的问题在于，比较句研究尚没有完全反映中文比较句的特征，尤其是概括比较属性（attribute, ATT）和比较结果（result, RES）之间多重关系的细颗粒度的模板。宋锐等（2009）以比较句中的介宾结构为主要研究对象，提出一些概括词性、句法位置的比较关系特征集。宋锐等也承认比较结果成分本身的复杂性使得目前的抽取方式还亟待进一步改善（2009:107）。可见，剖析其他要素与比较结果参杂交织而形成的复杂结构对于比较句比较关系的抽取起着重要的作用。因此，本文将对比较句关键要素的种类和数量进行扩充，并在此基础上标注这些关键要素在句子中所表现出来的语义角色类型，从而为下一步寻找出关键要素的序列模式（sequential pattern）奠定基础。我们希望这种从语言学角度出发，通过对大规模网络文本中真实句子的细致分析，最终可以为信息抽取中的比较句关键要素识别和提取提供一定的帮助。

2 汉语比较句

关于汉语比较句的研究很早就已经开始了，《马氏文通》（1898）把比较句分为“平比”、“差比”和“极比”三大类别，这三种类别的含义分别对应于英语中的原级、比较级和最高级的意义。吕叔湘（1982：352）把比较论述为两件事物之间同中见异，或异中见同的过程。他指出，比较两件东西的高下，高者对下者说是“胜过”。张晨（2013：112）认为，在形式上，以比较特

* 基金项目：国家自然科学基金项目（61375074，61371129）

[1] COAE 2012 任务 2.2（比较句关键要素提取）与其他任务相比显示较低的成绩（最佳 F 值仅为 0.35）。

征词的出现与否为准，汉语比较句可以划分为显性比较句与隐性比较句。隐性比较句指的是虽然句子中体现出比较意义，但是“比、最、更”等比较标志并没出现在句子里。

根据已有的研究成果，本文将“比”字句划分为四种类型，如下表 1 所示。

类型	语义	例句
胜过	在某种属性上比较主体的程度高于比较客体（基准）	安全、操控性都**比**富康、嘉年华好！ 高速 120 左右胎噪**比**发动机声音要大。
等同	两者之间的相同	335LI 配置**不比** 535 差。 做工**不比**国外货差。
不及	在某种属性上比较主体的程度低于比较客体	性价比高，音质**比不上**联想的，但也算不错的了。 诺基亚机子还行，但还是**比不过**摩托。
引入	把比较两项引入比较	**与**现有车型**相比**，车高降低了 100mm。 **比起**全屏幕的，我感觉键盘的更好。

表 1—“比”字句的类型

本文的主要分析对象是表达“胜过”义的显性比较句，即包含介词“比”的单句形式。因为本研究侧重于挖掘在“比”字句中关键要素的序列模式，而带从句的复句形式往往引入其他次要要素以及主观焦点，给普遍的序列带来不少的干扰。所以本研究中，不考虑以“与……对比/相比”为从句的复句形式（表 1 中的“引入”类）的比较句。本文从COAE 2013 所提供的电子、汽车两个领域的 1,000 句的比较句训练集中，挑出“胜过”义的 460 多句作为分析数据 [2]（其中汽车领域共 202 句，电子领域共 262 句），总结出这些“比”字句中关键要素的排序以及要素之间的比较关系。

3 “比”字句的关键要素及其定义

从工程角度上，“比”字句可以定义为一个由比较主体（SUB）、比较客体（OBJ）、比较属性（ATT）、比较结果（RES）和比较标记（BI）组成的五元组（quintuple）：

C = < SUB, BI, OBJ, ATT, RES >

目前大多数比较句识别模型在此组合的基础上对“比”字句进行分析。然而，我们认为这一定义对关键要素的识别和关系抽取不够精密，会错失相当多的语言学的启发性特征，尤其是比较属性（ATT）和比较结果（RES）之间所隐含的多种配位方式。因此，我们对上面的这一组合进行细化而补充，从而得出以下的七元组（septuple）：

C = < SUB, BI, OBJ, ITM, DIM, RES, EXT >

每个元素的定义如下：

比较主体（SUB）：比较两项中先出现的成分，比较结果（RES）这个属性值的持有者，一般充当句子的主语。

比较词（BI）：“比”字句标记，本文中为介词“比”。

比较客体（OBJ）：比较两项中后出现的成分，经常充任“比”字所引导的介宾结构的宾语，又称为比较基准。

比较项目（ITM）：语义上泛指 SUB 与 OBJ 共有的部件，比较结果（RES）的次要论元。一般与 SUB 和 OBJ 结合成名词短语，充任主语或由“比”引导的介宾结构的宾语。

比较维度（DIM）：语义上泛指SUB与OBJ共有的属性（property），比较结果（RES）的主要论元 [3]。一般与SUB和OBJ结合成名词短语，充任主语或由“比”引导的介宾结构的宾语。

比较结果（RES）：表达比较两项之间的差异，而且是比较维度（DIM）的属性值。语法上由形容词来实现，充当谓语。

比较量幅（EXT）：表达比较两项之间差异的幅度，一般和比较结果（RES）结合成述补结

[2] 通过“比”字匹配方式过滤（排除）非“比”字句（如包含“没有、不如、差不多”等比较词的句子），之后人工挑选出表“胜过”义的 460 多个“比”字句。

[3] 在语义网络中，类（class）、属性（property）和属性值（property value）构成基本的语义单位。Fellbaum（1998：40）举“知更鸟（robin）”的例子阐述这三种要素。名词“知更鸟”是属性尺寸（size）和颜色(color)的持有者，也能理解为这两种属性的论元，比如 SIZE(robin)=small、COLOR(robin) = red，而形容词“小（small）和红色（red）”分别是对于这两种属性的值。

构。

我们以下面这句为例来展示对句子所进行的上述七个元素的标注工作。

（1）新飞度车身结构的刚度比前代提高了164%。

新飞度	车身结构	刚度	比	前代	提高	164%
SUB	ITM	DIM	BI	OBJ	RES	EXT

4 “比”字句关键要素序列特征分析

“比”字句表达“胜过”概念具有十分复杂的模式。在实际文本中，上述七种要素纵横交错在一起实现为十分灵活的组配模式。现有的自动识别模型已对这些变化无穷的“比”字句进行了归纳，并建成了一些模式库。比如，宋锐等构建了包含106条比较模式的差比模式库（2009:104）；侯明牛等（2013）归纳出否定比较句的若干句型。另外，陈珺等（2005）从教学大纲里的比较句语法项目中选取了具有代表性的20多种比较句型。

虽然比较模式表面上看起来五花八门，但我们相信实际上存在一些把它们贯串起来的规则或倾向性。我们针对电子和汽车这两个领域中的460多个“比”字句，按照上述七元组的基本架构进行了人工标注。经过加工与简化的步骤后，我们得出以下6种（R1-R6）基本序列规则，分列如下。

4.1 比较属性（DIM）与比较结果（RES）的排序

R1：比较结果（RES）不能出现于比较维度（DIM）前面。并且，比较结果（RES）必须在句中出现。例如：

(2) NOKIA的信号比我另外一个两千多的手机信号还好。

NOKIA	信号	比	两千多的手机	信号	好
SUB	DIM	BI	OBJ	DIM	RES

上述例句中，比较维度（DIM）先于比较结果（RES）出现。比较结果（RES）的含义则是针对某种属性（property）的具体值（property value）。从这个角度来看，比较维度（DIM）即是比较结果（RES）所要描述的属性。因此，描述的属性必须先被确定并出现，然后才能对它进行陈述，即比较结果（RES）随后出现，结果产生DIM-RES的常规语序。

但是，实际语料中也存在一些RES-DIM逆序的情况。条件是：句子中出现浮现动词（Emergent Verb），比如下面三句中的动词“提供、具有、多出”。

(3) xD卡比市面上的其他存储介质提供了更高速的连接速度与更小的体积。

xD卡	比	其他存储介质	高速	连接速度	小	体积
SUB	BI	OBJ	RES	DIM	RES	DIM

(4) LTPS屏幕比TFT屏幕具有更高的色彩饱和度和亮度。

LTPS	屏幕	比	TFT	屏幕	更高	色彩饱和度、亮度
SUB	ITM	BI	OBJ	ITM	RES	DIM

(5) 无优点，就算有，也是比卡片机多出个手动模式。

比	卡片机	多出	手动模式（数目）
BI	OBJ	RES	ITM（DIM）

在这些RES-DIM逆序，即比较结果（RES）前置的非常规句子中大部分情况是有浮现动词充当“比”字句的谓语。这又可以细分为两种情况：①比较结果（RES）充当比较维度（DIM）的定语，如(3)、(4)所示；②比较结果（RES）携带比较维度（DIM）宾语，如(5)。

4.2 比较属性相对于比较词的位置

R2：比较属性（ITM和DIM）一般在“比”前出现

在“比”字句中，比较属性（ITM与DIM）是比较的核心对象，即比较点。因此，它们容易作为“比”字句的话题，提升到句首位置。尤其是ITM，它经过话题化前置之后，能替代比较主体（SUB）。具体统计结果如下表2、3所示。

出现项：ITM	“比”前	“比”后	前后	ITM 总计
CAR	33	1	2	36
DIGITAL	25	5	7	37

表 2—比较项目（ITM）出现位置分布

出现项：DIM	“比”前	“比”后	前后	DIM 总计
CAR	99	20	5	124
DIGITAL	114	24	9	147

表 3—比较维度（DIM）出现位置分布

ITM 一般在“比”前面出现（33/36, 88%；25/37, 68%），DIM 也倾向于出现在“比”前面（99/124，80%；114/147，78%）。

出现这种现象，我们认为原因是：在“比”字句中，比较属性 DIM、ITM 的描述对象主要是比较主体（SUB），而不是比较客体（OBJ）。ITM 和 DIM 这两种范畴是比较项（SUB 和 OBJ）都具有的共同属性。“比”字句（差比）核心功能是从其共同属性中显露出比较差异，尤其是关于比较主体（SUB）的比较属性。因此，比较点落实在比较主体（SUB）的部件（ITM）或属性（DIM）上，而不会落实在比较的参照标准，即 OBJ 上。因此，序列模式 A 能够成立，而像 B 这样的序列绝不能出现。

A.（SUB） ITM DIM BI OBJ RES

(6) 显然，Xg 在电池的续航能力方面比 Xt 有所提高。

Xg 电池 续航能力 比 Xt 提高

SUB ITM DIM BI OBJ RES

(7) 键盘手感比 E71 好，耳机插孔是 3.5mm，性能很快，价格合适。

键盘 手感 比 E71 好

ITM DIM BI OBJ RES

*B. (SUB) BI OBJ ITM DIM RES

(8) 从使用的过程中来看，比前两代机型的屏幕不管在色彩、亮度方面都有非常大的提高。

比 前两代机型 屏幕 色彩、亮度 提高

BI OBJ ITM DIM RES

在实际文本中能归入到序列 B 的情况，我们只发现了 1 句（例 8）。该句中的“在……方面”将 DIM 突出加以强调，使得 DIM 句法位置较为灵活，可以出现在 OBJ 之后。如果没有“在……方面”，句子的可接受度将大大降低，比如：

(8') ？从使用的过程中来看，比前两代机型的屏幕色彩、亮度都有非常大的提高。

此外，ITM和DIM对于SUB的语义偏向也反映在关键要素之间的关系规则（association rule）上。我们从 464 句表示胜过义的“比”字句抽取了由关键要素序列构成的 469 个事务（transaction）[4]。例如，上述例句（6）的事务为 t_i={SUB, ITM, DIM, BI, OBJ, RES}。

我们以 469 个事务为分析对象，利用Apriori算法挖掘出包含DIM、ITM的频繁项集合（frequent itemset）以及其关系规则（association rule[5]）。其中，与DIM共现的频繁项（1-item）按置信度[6]降序排列为：

RES（93%）SUB（62）EXT（37）OBJ（13）ITM（6）：CAR

RES（92%）EXT（37）SUB（21）OBJ（14）ITM（6）：DIGITAL

[4] 有些例句是由两个以上的“比”字句分句构成的，我们把分句分为独立的事务。

[5] 关系规则（association rule）：项目的集合 $I = \{i_1, i_2, \ldots, i_m\}$，事务 $T = (t_1, t_2, \ldots, t_n)$, $t_i \subseteq I$。关系规则可视为 $X \rightarrow Y$（$X \subset I$, $Y \subset I$, $X \cap Y = \varnothing$），其中 X（或 Y）叫做项集（itemset）。（Bing Liu 2011:18）

[6] 置信度（confidence）是在关系规则中常用来衡量关系规则强度的指标。一个关系规则 $X \rightarrow Y$ 的置信度是指既包含了 X 又包含了 Y 的事务（transaction）的数量占所有包含了 X 的事务的百分比。（Bing Liu 2011:18）

与 ITM 共现的频繁项（1-item）按置信度降序排列为：

RES（92%）SUB（62）EXT（38）DIM（21）OBJ（3）：CAR

RES（84%）SUB（25）DIM（20）EXT（18）OBJ（11）：DIGITAL

与两种关键要素 DIM 和 ITM 共现的频繁项排序都显示一致的结果，即比较主体（SUB）的置信度远远高于比较客体（OBJ）。这说明了与比较客体（OBJ）相比，比较主体（SUB）与比较属性（DIM、ITM）共现的频率更高，意味着 DIM 或 ITM 的描述对象大部分是比较主体（SUB），而不是比较客体（OBJ）。

4.3 比较属性（ITM 和 DIM）内部序列

R3： 如果句子中比较属性的两种范畴（ITM 和 DIM）同时出现，一般 ITM 出现于 DIM 前面，而 DIM 不能出现于 ITM 前面，例如：

(9) 比较便宜，外观不错，喇叭声音也比宏基以前的机型大了不少。

喇叭	声音	比	宏基以前的机型	大	不少
ITM	DIM	BI	OBJ	RES	EXT

在（9）中 ITM 出现于 DIM 前面，而 DIM 不能出现于 ITM 前面。与 R1 的机制一脉相承。DIM 是 ITM 的属性（property）。属性是相对于实体（Entity），即 ITM 而言的。如果没有和自己能够对应上的实体，即 ITM，属性（DIM）是站不住的。因此，在语序上先要确定实体（ITM），之后对该实体的属性（DIM）才能进行补充描述（R3）。

R4：DIM 与 ITM 不能由“比”来隔开。因此，以下序列是不允许的。

A. * DIM * BI * ITM * （*表示任何成分）

B. * ITM * BI * DIM *

根据 R4，DIM 与 ITM 在同一个句子中一起出现时，要么都在“比”前面，要么都在“比”后面。可是后者的情况被 R2 排除，因此，DIM 和 ITM 只能整体在“比”前出现。例如：

(10) 急加速时发动机的噪声比福克斯强多了。

发动机 噪声	比	福克斯	强	多
ITM DIM	BI	OBJ	RES	EXT

4.4 比较属性（ITM 和 DIM）的分布规则

R5：ITM 和 DIM 的分布大体上是互补的。

通过对语料的整理，我们发现：DIM 在语义上是自足的，而 ITM 不是自足的。我们可以再引入 4.2 的关系规则（association rule）来对此加以验证：

与 DIM 共现的频繁项（1-item）按置信度降序排列为：

RES（93%）SUB（62）EXT（37）OBJ（13）ITM（6）：CAR

RES（92%）EXT（37）SUB（21）OBJ（14）ITM（6）：DIGITAL

与 ITM 共现的频繁项（1-item）按置信度降序排列为：

RES（92%）SUB（62）EXT（38）DIM（21）OBJ（3）：CAR

RES（84%）SUB（25）DIM（20）EXT（18）OBJ（11）：DIGITAL

上述关系规则显示，规则 ITM→DIM 的置信度是 20%，而规则 DIM→ITM 的置信度只有 6%。我们据此可以把它理解为比较项目（ITM）对比较维度（DIM）的依赖性比 DIM 对 ITM 的依赖性更强。从中推导出在“比”字句中，比较项目（ITM）是不能单独出现的，而是需要通过比较维度（DIM）作为媒介才能实现。

这种 ITM 与 DIM 之间的置信度不对称现象（20%、6%）说明，这两个成分的出现与否并非互相独立的事件，而是存在某种相互关系。如下面的表 4、5 所示。

	ITM 出现（句）	ITM 未出现（句）
DIM 出现（句）	9	116
DIM 未出现（句）	28	61

表 4—比较项目（ITM）与比较维度（DIM）出现与否的交叉分布（汽车领域）

	ITM 出现（句）	ITM 未出现（句）
DIM 出现（句）	10	137
DIM 未出现（句）	27	89

表 5—比较项目（ITM）与比较维度（DIM）出现与否的交叉分布（电子商品领域）

通过汽车、电子商品领域的卡方检验（chi-squared test），我们发现了ITM与DIM之间的确存在比较强的相关关系[7]。换句话说，ITM与DIM这两种要素的出现与否存在相互关联，在表 4、5 里，左下右上的出现频率尤其显著。具体而言，ITM在“比”字句中出现时，DIM倾向于不出现，反之亦然。那为何ITM和DIM具有这种较强的互补分布呢？这是我们把比较属性（ATT）分成ITM和DIM两类的自然结果。比较属性（ATT）是说话人要对比的比较焦点，最好只有一个，因为两个或两个以上的比较焦点，容易造成认知上的负担。比如，在文本中会经常出现下面A、B两种情况，能够明显地表现出这一倾向性。

A. ITM 出现，但 DIM 不出现的情况

SUB	**ITM（DIM）**	**比**	**OBJ**	**RES**	**（EXT）**
	车身（重量）[default]	比	老款伊兰	轻	
逍客	内部空间（体积）[default]	比	CRV	小	很多
	底盘（强度）[default]	比	雅力士	硬	
	轴距（长短）[default]	比	凯悦	长	10mm
	硒鼓（价格）[default]	比	白色	便宜	

在上述例句中，比较焦点落实在 ITM 上。DIM 都是 RES 的缺省属性（default property），因此不用再重复提示 DIM，可以直接把它省略。

B. DIM 出现，但 ITM 不出现的情况： 比较焦点落实在 DIM 上，因此 ITM 没有浮现。

SUB	**DIM**	**比**	**OBJ**	**RES**	**（EXT）**
夏利	性能	比	吉利车	好	
FIT	外观	比	同类车	大气	
LEXUS	起步	比	我	快	很多
广州本田	定价	比	同级别的车	低	5 万
新天籁	油耗	比	上一代	降低	10%

C. ITM、DIM 同时出现的情况

(SUB)	**ITM**	**DIM**	**比**	**OBJ**	**RES**	**（EXT）**
	发动机	噪声（强度）[default]	比	福克斯	强	多
	屏幕	分辨率（高度）[default]	比	同价位小米	高	
Thinkpad	屏幕	可视度（大小）[default]	比	以前的	大	
Xg	电池	续航能力（质量）[default]	比	Xt	提高	
宝来	刹车系统	反应（速度）[default]	比	307	慢	

情况 C 是 R5 的反例，但如表 4、5 所示，这种情况并不多（4%）。在情况 C 中，虽然比较项目（ITM）在句子中已出现，但是这个事实不能保证比较维度（DIM）一定省略而不出现。这是因为从比较结果（RES）的意义上不能推导出 DIM，因此为了避免模糊性，比较维度（DIM）就很有可能就浮现。置于 DIM 的省略条件，我们在 4.5 节继续探讨。

[7] 在汽车、电子商品领域的卡方值（χ^2）分别为 21.39 和 14.55，大于临界值 3.84（df=1, α=0.05）。

4.5 成分省略规则

R6：“比”字句中，比较词（BI）、比较客体（OBJ）和比较结果（RES）三种要素强制性出现。

一旦比较属性（DIM 和 ITM）在“比”字句中出现，比较焦点从原有的 SUB 和 OBJ 转移到比较属性 DIM 和 ITM 上，成为“比”字句的核心比较焦点。这时，关键要素的省略应该考量以下三点。

R6-1：比较结果（RES）的语义指向是比较维度（DIM），不是比较项目（ITM）。因此，原则上 DIM 不能省略。

与 RES 共现的频繁项（1-item）按置信度降序排列为：

SUB（71%）DIM（57）EXT（39）OBJ（37）ITM（17）

根据上述排列，与 RES 共现概率较高的关键要素为比较主体（SUB）和比较维度（DIM）。与这两个关键要素相比，ITM 的置信度非常低（17%），表示它不是实现比较结果（RES）的必要条件。与 ITM 不同，SUB 和 DIM 两个成分在述谓结构上正好是属性值（property value，即 RES）的两个论元，即实体（Entity，即 SUB）和属性（Property，即 DIM）。从述谓结构的角度看，比较结果（RES）要求以 DIM 和 SUB 的出现为自己的实现条件。因此，DIM 和 SUB 必须在比较结果（RES）前出现。

R6-2：比较结果（RES）决定比较维度（DIM）是否省略

在大约 40%的句子中，DIM 没有实现，这是因为我们一般能猜测到 RES 所默认（default）的属性（Property），即 DIM。

从上述 4.4 节中的三种情况中，我们发现了一个较为重要的规则，即比较维度（DIM）的省略与否与比较结果（RES）之间存在一定的关系。DIM 省略而不出现时（4.4 节 A），DIM 一般都是比较结果（RES）的缺省属性（default property），如：“车身（重量）[default] 轻”。然而，DIM 不省略而出现时（4.4 节 B 和 C），DIM 不是 RES 的缺省属性。因此，DIM 一般不能从 RES 的含义推导出来：“这款车比那款车维修费用高”中的比较维度（DIM）“维修费用”不能省略，因为比较结果（RES）“高”的缺省属性是“高度”。主要比较结果（RES）及其缺省属性如以下表 6 所示：

比较结果（RES）	缺省属性	非缺省属性（句中 DIM 显现）
高/低	高度	油耗|性价比|分辨率|最大功率、最大扭矩、最高车速|清晰度|耗材通用性|速度、精度|像素|性能|配置|维修费用|价格|风阻、车身尺寸、车重
多/少	数量	油耗|性价比|功率|油费|内饰|噪声|调节|功能|价格
大/小	体积	最大扭矩|可视范围|排量|噪声|后排空间|声音|功率|震动|车重|音量|噪音|可视角|最大扭距|马力
快/慢	速度	运行|打印速度|反应|系统启动、运行速度|起步
重/轻	重量	车重|整备质量|抖动|质量
贵/便宜	价格	养路费|保养
提升/提高	水平、质量	速度|刚度|噪音|做工|强度|续航能力|色彩、亮度|输出功率
好/差	品质	声效|性价比|画质|感觉|打印质量|做工|性能|安全性能|效果|开|手感|安全性|分辨率|待机|做工|打印速度|音质|样式|减震性|优化方面|声音|显示效果|散热|信号|手感、工艺、屏幕色彩|信号、通话|刹车|拍|配置|外观|加速性能|稳定性、加速、隔音|彩打效果|照相|降温效果|打字|铃音效果|其他性能|安全系数|内饰|排量、座位、尺寸、装饰|清晰度|操作系统|细节方面|档次|配置|整体性能
强	品质	内饰|性价比|地盘悬挂、隔音、发动机动静|中段的再加速|反映|待机|做工|曝光量|噪声|开放性|续航能力|待机时间|动力|感光度|品牌效应和售后服务
好看/难看	外观	系统界面|车款

耐用	耐久性	无（属性-属性值一体化）
好用	使用性	无（属性-属性值一体化）
省油/费油	油耗	无（属性-属性值一体化）

表 6—比较结果（RES）的缺省属性及非缺省属性（DIM）

表 6 说明，当比较结果（RES）的缺省属性和句子中的DIM不一致时，DIM则强制性出现；缺省属性和句中DIM一致时，DIM非强制性出现。为了易于理解，我们认为表 6 中的非缺省属性（DIM）是比较结果语义溢出[8]的结果，例如：

(11) 我听网友们评论老款三缸 376 夏利机比四缸机马力还要大些，你感觉怎么样?

老款三缸 376 夏利机　比　四缸机　马力　大些

SUB　BI　OBJ　DIM　RES

若 DIM“马力”未出现，例（11）的比较结果（RES）“大”所描述的属性不明确，因此“大”所隐含的属性“马力”从比较结果溢出而显现了。另外，作为一种特殊情况，有些 RES 的属性-属性值一体化结构（耐用、好用、省油 等）已确定其缺省属性，因此 DIM 强制性省略（表 6）。

比较结果（RES）	意义	缺省值	句中出现的 DIM
扎实	结实	?	做工
实在	真实；不虚假	?	价格
充沛	充足而旺盛	?	动力
省	节约	?	耗油|修理|价格|油
可靠	可以信赖依靠	?	质量
自然	不勉强；不局促；不呆板	?	色彩
专业	技术或知识掌握得很透彻	?	系统
大气	大而霸气	?	外观|外形
细致	精细周密	?	做工
全面	各个方面的总和	?	安全系统
高档	质量好，价格较高的（商品）	?	内饰
清晰	清楚	?	语音|像素

表 7— 难以决定缺省属性的比较结果（RES）

另外，有的比较结果（RES）本身不具有或无法指定其缺省属性（表 7），因此不能推导出相对应的比较维度（DIM）。这时，仅仅通过比较结果（RES）的意义不能预测比较维度（DIM）的具体属性。如：

(12) 夏利比吉利修理上省一点，因为夏利车是个汽修店都能修。

夏利　比　吉利　修理　省　一点

SUB　BI　OBJ　DIM　RES　EXT

(13) M6 的电子主动安全系统比力狮全面。

M6　电子主动安全系统　比　力狮　全面

SUB　DIM　BI　OBJ　RES

例（12）、（13）的比较结果“省、全面”无法指定其缺省属性。这时我们不能预测它们的 DIM 究竟指的是什么。与表 6 中的情况不同，表 7 中的比较结果（RES）一般不能省略其比较维度（DIM），即 DIM 强制性出现。

R6-3：比较主体（SUB）和比较维度（DIM）同时省略条件

虽然 SUB 和 DIM 同时省略的情况数量很少，但“SUB BI OBJ RES”能进一步省略比较主体（SUB）。比如：

[8] 关于语义溢出（semantic overflow）的详细内容，见袁毓林（2012）。

A. BI OBJ RES

（14）比一般的豪车还要费油,比 Q7 都多,只能归咎于技术不成熟,油耗控制在 10 左右,就很完美了。

比 一般的豪车 费油

BI OBJ RES

（15）出差日本用来暂时通话，比同档次的三星手机好用。

比 同档次的三星手机 好用

BI OBJ RES

A 这样的超短形式之所以被允许是因为虽然句子中省略了 RES 的述谓结构上的论元 SUB 和 DIM，但是还能从 RES 和 OBJ 推导出省略了的成分（DIM 和 SUB）。对于 DIM 而言，有些 DIM 作为缺省值被蕴含在 RES 的意义里面（表 6，属性-属性值一体化），DIM 不出现却还能推理出来。比如，（14）的比较结果“费油”默认它所评价的属性是“油耗”。对于 SUB 而言，虽然 SUB 被省略了，但从比较客体（OBJ）中还是能够推导出 SUB。这是因为 SUB 是与 OBJ 属于同一范畴的实体（Entity），比如：（15）的 OBJ“同档次的三星手机”代表某种品牌和商品。我们能够从中推导出和它相对应的 SUB，比如“同档次的联想手机”。

*B. BI ITM RES （不合格的省略句型）

在“比”字句中，B 这种序列基本上无法实现，这是因为比较项目（ITM）的语义具有不自足性。在序列 B 中，ITM 本身并不能推导出任何关于 SUB 的信息，因为 ITM 不像 A“BI OBJ RES”序列中的 OBJ 那样与 SUB 属于同一范畴的成分。ITM 只有 SUB 或 OBJ 出现的情况下才具有所指的对象。简而言之，ITM 的意义是不自足的。因此，序列 B“BI ITM RES”是非常规语序，无法实现。

5. 序列的合格性判断

我们以 469 个事务为分析对象，利用PrefixSpan算法挖掘出常见序列模式（sequential pattern[9]）。其次，根据在第 4 章所提出的六种规则，对实际文本中的序列模式进行合格性检查和特征归纳。序列模式按支持度（support）降序分列如下[10]。

（1）DIM BI OBJ RES （EXT） #SUP: 105(22.4%)

够用，待机时间比现在智能机长，老人用还算合适。

待机时间 比 智能机 长

DIM BI OBJ RES

RULE 2：比较属性一般在“比”前出现，当作话题。

RULE 4：DIM 与 ITM 不能由“比”隔开，因此“比”后不出现 ITM。

RULE 5：DIM 出现了，因此 ITM 不出现概率比较高。

RULE 6-2：由于 DIM“待机时间”不是 RES“长”的缺省属性（长度），因此DIM 出现。

（2）SUB BI OBJ RES (EXT) #SUP: 81 (17.3%)

买宝来都比买这款车合算些。

宝来 比 这款车 合算

SUB BI OBJ RES

RULE 6-2：由于 RES“合算”蕴含其缺省属性（价格），DIM 能省略。

（3）SUB DIM BI OBJ RES （EXT） #SUP: 79 (16.8%)

领驭的后门进入空间比老款帕萨特大得多。

领驭 后门进入空间 比 老款帕萨特 大 得多

SUB DIM BI OBJ RES EXT

[9] 序列模式(sequence pattern) 是指项目的集合 $I=\{i_1,i_2,...i_m\}$中序列(sequence) S = $<a_1,a_2,...a_r>$，a_i 属于项集 X, $X \subseteq I$. （Bing Liu，2011:42）

[10] 序列模式的支持度(support)是指给予的序列 S 数量占所有包含了其序列的事务(transaction)的百分比。此外，由于受篇幅限制，本文只展示常见的“比”字句的序列模式，而未展示非常见的“比”字句序列模式。

RULE 1：DIM 先于 RES 出现。
RULE 2：DIM 在“比”前面出现，因为 DIM 的指向是 SUB，不是 OBJ。
RULE 5：DIM 出现，则 ITM 没出现。
RULE 6-2：DIM“空间”不是 RES“大”的缺省属性（体积），因此属性 DIM 出现。

（4）BI OBJ RES (EXT) #SUP: 65(13.9%)
比安卓手机省电多了，外放铃声也大，是正品行货质量有保证，价位还算可以。
比 安卓手机 省电 多了
BI OBJ RES EXT
RULE 6-2）由于 RES“省电”蕴含其属性（电耗），所以 DIM 能省略。
RULE 6-3）OBJ 存在，能推导出省略了的同一范畴的 SUB（如，苹果手机）。

（5）SUB ITM BI OBJ RES （EXT） #SUP: 26(5.5%)
最喜欢的就是这车子的底盘比一般的同级车要高一些，过个减速带什么的感觉很好，很稳。
这车子 底盘 比 一般的同级车 高 一些
SUB ITM BI OBJ RES EXT
RULE 2：在“比”前仅出现 ITM，成为了比字句的话题（比较焦点）。
RULE 4：由于 ITM 在“比”前出现，其他比较属性（DIM）不能在“比”后出现
RULE 5：由于 ITM 已出现，因此 DIM 不出现的概率较高
RULE 6-2：句中 RES“高”所描述的属性是缺省属性“高度”。因此，DIM（高度）被省略。

（6）BI OBJ DIM RES (EXT) #SUP: 21(4.5%)
感觉比以前在京东买的 e430 的做工好一点，以前的面板和做工好差哦。
比 e430 做工 好 一点
BI OBJ DIM RES EXT
RULE 1：DIM 先于 RES 出现。
RULE 4：由于 DIM 在“比”后出现，因此 ITM 不在“比”前出现。
RULE 5：DIM 出现了，因此 ITM 会出现的概率较低。
RULE 6-3：比较客体（OBJ）是商品名，从中能够推导出 SUB 是属于同一范畴的商品名。

（7）ITM BI OBJ RES （EXT） #SUP: 20(4.3%)
底盘比雅力士要硬，但更粘地面一些，指向性强。
底盘 比 雅力士 硬
ITM BI OBJ RES
RULE 2：比较属性（DIM 和 ITM）一般在“比”前出现，充当话题。
RULE 4：DIM 与 ITM 不能由“比”隔开，因此“比”后不出现 DIM。
RULE 5：ITM 已出现，因此 DIM 不出现概率比较高。
RULE 6-2：由于 RES“硬”的属性“硬度”是缺省属性，因此 DIM 省略。

（8）SUB ITM DIM 比 OBJ RES （EXT） #SUP: 5 (1.1%)
新飞度车身结构的刚度比前代提高了 164%。
新飞度 车身结构 刚度 比 前代 提高 164%
SUB ITM DIM BI OBJ RES EXT
RULE 1：DIM 在 RES 前出现。
RULE 2：ITM 和 DIM 一般在“比”前出现。
RULE 3：ITM 在 DIM 前出现。
RULE 4：ITM 与 DIM 不能由“比”来隔开
RULE 6-2：RES“提高”的缺省属性不是“刚度”，因此 DIM（刚度）不能省略。
RULE 6-3：因为从 OBJ“前代”不能推导出 SUB“新飞度”，SUB 不能省略。

6. 总结和展望

本文针对句子级比较关键要素挖掘问题，尤其是汉语“比”字句关键要素识别，重新划分了“比”字句关键要素的范畴。通过研究，我们发现“比”字句中的比较项并不局限于比较主体（SUB）和比较客体（OBJ）。在实际文本中，很多情况下比较项是属于不同范畴、不同级别的两个成分，比如，“做工，外观比它们漂亮”中的比较项是属于不同范畴的 DIM（属性）与 OBJ（实体）。这种比较项之间的异质性增加了比较结果（RES）的语义倾向性（Polarity）判断难度。尤其是“大/小、高/低、多/少”等情感模糊词充当比较结果（RES）时，找出它们相对应的比较维度（DIM）是对整个句子做出正确的极性判断的至关重要的问题。为了解决这些问题，本文在目前通用的五种基本要素分类基础上，补充了比较项目（ITM）、比较维度（DIM）和比较量幅（EXT）三种要素，提出了由七种关键要素构成的“比”字句的分析架构。然后，利用一些序列模式挖掘算法抽取“比”字句的多种模式，从中抽出共六种汉语“比”字句序列规则，并揭示这些规则背后的语言学动因。研究的后续工作包括搜集更多的语料、扩充模式规则并概括“比”字句以外的比较模式。另一方面，构建和比较结果（RES）词共现的比较项目（ITM）、比较维度（DIM）词汇库也是亟待研究的课题。我们期望利用该词汇库信息能抽取比较属性（ITM 和 DIM）以及比较主体（SUB）。本文的研究结果对“比”字句关键要素抽取任务的实际应用与评测具有重要的参考价值，也将为后续的“比”字句关键要素的自动提取工作奠定基础。

参考文献

[1] 陈珺,周小兵. 比较句语法项目的选取和排序[J], 语言教学与研究，2005，（2）：22-32.
[2] 耿直. 基于语料库的比较句式“跟、有、比”的描写与分析[D].北京：北京大学，2012：111-164.
[3] 侯明午，周红照，程南昌 等. 汉语否定比较句句型研究及在工程中的应用[R].第五届中文倾向性分析评测会议，太原，2013：88-96.
[4] 黄小江，万小军，杨建武 等. 汉语比较句识别研究[J]. 中文信息学报, 2008,(5):30-38.
[5] 贾玉祥，昝红英，范明. ZZU_NLP：汉语比较句识别第[R].五届中文倾向性分析评测会议，太原，2013：102-105.
[6] 李临定. 现代汉语句型[M]. 北京：商务印书馆，1986：285-301.
[7] 李岩，徐蔚然，陈光. PRIS_COAE COAE2013 评测报告[R].第五届中文倾向性分析评测会议，太原，2013：53-69.
[8] 刘月华 等. 实用现代汉语语法. 北京：商务印书馆，2001：833-854.
[9] 吕叔湘. 中国文法要略[M]. 北京：商务印书馆,1982：352-370.
[10] 马建忠. 马氏文通[M]. 上海：商务印书馆, 1898.
[11] 廖祥文，许洪波，孙乐，等. 第三届中文倾向性分析评测（COAE2011）语料的构建与分析[J]. 中文信息学报，2013，（1）：56-63.
[12] 宋锐，林鸿飞，常富洋. 中文比较句识别及比较关系抽取[J], 中文信息学报，2009，（3）：103-122.
[13] 谭松波，王素格，廖祥文 等. 第五届中文倾向性分析评测（COAE2013）总结报告[R].第五届中文倾向性分析评测会议，太原，2013：5-33.
[14] 谭松波，王素格，廖祥文 等. 第五届中文倾向性分析评测（COAE2013）标注数据（Task2）[DB]. 2013. http://ccir2013.sxu.edu.cn/COAE.aspx.
[15] 魏现辉，任巨伟，何文译 等. DUTIR：中文短文本倾向性分析及要素抽取方法研究[R].第五届中文倾向性分析评测会议，太原，2013：116-129.
[16] 王素格，王凤霞，宋雅，等. 基于序列模式的汉语比较句识别方法[J]. 山西大学学报，2013，（2）：172-179.
[17] 许国萍. “比”字句研究综述[J],汉语学习，1996，（6）：28-31.
[18] 袁毓林. 动词内隐性否定的语义层次和溢出条件[J]. 中国语文. 2012,(2)：99-113.
[19] 袁毓林. 基于生成词库论和论元结构理论的语义知识体系研究[J]. 中文信息学报. 2013，（6）：23-30.
[20] 袁毓林. 形容词的语义特征和句式特点之间的关系[J]. 汉藏语学报. 2013，（7）：147-166.
[21] 张晨，马冲，刘全超 等. 基于多特征融合的中文比较句识别算法[J], 中文信息学报，2013，（6）：111-116.
[22] 中国社会科学院语言研究所. 现代汉语词典（第 6 版）[M]. 北京：商务印书馆, 2012.
[23] Bing Liu. Web Data Mining: Exploring Hyperlinks ,Contents ,and Usage Data[M]. Springer , 2011.

[24] Christiane Fellbaum. WordNet: An Electronic Lexical Database[M]. Massachusetts: MIT Press, 1998: 23-43.
[25] Fournier-Viger, P., Gomariz, A., Soltani, A., et al. SPMF: Open-Source Data Mining Platform[CP]. 2014. http://www.philippe-fournier-viger.com/spmf/.
[26] John Lyons. Semantics, volume 1[M]. Cambridge: Cambridge University Press, 1977: 270-280.
[27] Peter Harrington. Machine Learning in Action[M]. Manning Publications Company, 2011:224-266.
[28] Yunfang Wu, Miaomiao Wen. Disambiguating Dynamic Sentiment Ambiguous Adjectives[R]. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, 2010: 1191-1199.