北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.030

基于句法语义规则系统的比较句自动识别

白林楠† 胡韧奋 刘智颖

北京师范大学中文信息处理研究所，北京 100875; † 通信作者, E-mail: linnanbai@126.com

摘要 针对汉语比较句的自动识别，提出一种基于句法语义规则的方法。比较标记和比较结果是识别比较句的关键因素，在此基础上归纳汉语比较句的类别，书写比较句识别规则，同时设计 4 个模型进行分类识别。实验结果表明，规则系统可以有效地实现汉语比较句的句法分析和自动识别，为比较关系的抽取打下良好的基础。

关键词 比较句；自动识别；句法语义规则

中图分类号 TP391

Recognition of Comparative Sentences Based on Syntactic and Semantic Rules-System

BAI Linnan†, HU Renfen, LIU Zhiying

Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875; †Corresponding author, E-mail: linnanbai@126.com

Abstract The authors propose a novel method to identify comparative sentences based on rules, and these rules contain syntactic and semantic features of comparative sentences. Comparative marks and comparative result words are significant elements to identify comparative sentences. Based on this, the authors conclude the categories and identification rules of comparative sentences. Four models are designed to respectively recognize every category. Experiments show that proposed method can gain satisfactory results in comparative parser and recognition, which lay good foundation for comparative relation extraction.

Key words comparative sentences; recognition; syntactic and semantic rules

汉语比较句是对外汉语教学中的一个重难点，在经典对外汉语教材中，比较句约占 10%。汉语比较句有丰富的句法形式，除了最典型的“比字句”，还有许多句法形式也可以表达比较的含义。“有”字就可以和“比”字一样表示比较，如“我游得不好，没有你游得好”；“于”作为一个类词缀，也可以表示比较的含义，如“在衣食住行四事之中，食的程度高于其余一切”。传统的对外汉语教学模式无法对比较句的教与学进行有针对性的指导，因此自然语言处理的技术急需应用到对外汉语教学中。随着自然语言处理技术的发展，句法分析器能够为对外汉语教学提供越来越多的帮助。

在自然语言处理领域，关于比较句的研究越来越成为自然语言处理学科中的一项热点。比较句的研究对于文本挖掘、情感分析等具有重要的价值。国内外对比较句的识别研究主要集中在比较句自动识别和比较关系的抽取上，通用的技术是利用 SVM (Support Vector Machine)和 CSR (Class Sequential Rules)算法进行比较句识别，利用 LSR (Label SequentialRules)算法进行比较关系抽取。以上都是用统计的方式研究比较句识别，目前利用规则对比较句的研究非常少。本文的研究主要依据概念层次网络(HNC)理论，在对比较句进行分类和分布研究的基础上，以规则的方式进行比较句句法分

国家高技术研究发展计划专项经费(2012AA011104)资助
收稿日期: 2014–06–29; 修回日期: 2014–10–13; 网络出版时间: 2014-12-01 09:26

析和自动识别。

1 相关工作

在理论语言学界，语言学家主要关注比较句的定义、分类以及比较句的语法、语义和语用上的理解。丁崇明[1]认为“比较句是出现比较项，表示比较语义的句子”。刘月华等[2]没有专门定义比较句，但通过论述“比较的方式”阐述了比较句的内容。尚平[3]系统总结了马建忠、黎锦熙、高名凯、吕叔湘、丁声树 5 位大家对比较句的分类结果，先辈们从语言描述的角度进行细致的研究分析，为比较句自动识别奠定了坚实的理论基础，但几位语言学大家研究重点在于比较的意义，因此扩大了比较句的外延。比如，吕叔湘先生根据比较程度把比较句划分为 10 个小类，但其中的“高下”实际上是一种选择，而非比较。车竞[4]认为“现代汉语比较句是指句子中含有比较词语或比较格式的句子”，在构成上，有比较主体、比较客体、比较词、比较点、比较值等成分，分为 17 类比较句式。

理论语言学家侧重于对比较句这一特殊句式的描写和分析，而对外汉语界的语言学家则更侧重于教学实际，他们对比较句的研究重点是可操作性强的比较句语法项目的选取和排序。丁崇明[1]把比较句分为差比和平比。差比有“比”字句、“没(有)”句、“不如”句、“(和)……不一样”句、“于”字比较句和无标记的比较句，平比句包括“和……一样”以及“有”字句。同时穷尽了每一类比较句下的详细的句型。刘月华[2]把“比较”分为两大类：一类是比较事物、性状的同异；一类是比较性质、程度的差别、高低。具体分类有：A 跟 B 一样、A 有 B 那么(这么)……；“比”字句、“不比”句、“没有”句、“不如”句。陈珺等[5]参照各大语法教学大纲，考察中外语料中比较句的使用频率和偏误，为对外汉语教学比较句语法点提供等级选择和排序，共有 20 项详细分类。

语言学家为比较句的信息处理提供了语言学基础，这些研究对于建立比较词知识库、总结比较句识别规则有很大帮助，但与自动识别还有很大的距离。在自然语言处理领域，比较句的研究越来越引起人们的注意。在英语方面，Nitin[6–7]等首先利用 SVM 和 CSR 算法对比较句进行识别，最高取得 84%的准确率和 83%的召回率；然后又利用 LSR 算法对比较元素进行抽取，取得较好的效果。在汉语方面，北京大学的黄小江等[8]在 Nitin 研究的基础上,利用 SVM 和 CSR 对汉语比较句进行识别，取得了较高的准确率,但是召回率不高，平均只有 70%左右。张辰[9]提出一种基于规则与统计相结合的方法并进行实验。黄高辉等[10]利用 CRF 算法对汉语比较句进行自动识别和比较关系的抽取，取得较好的结果。

尽管比较句的识别取得了较好的结果，但目前比较句识别和分析还没有很好的应用，更没有关注对外汉语学界中比较句的识别与应用。另外，之前的研究采用的语料多是新闻语料，比较句的分布没有对外汉语教材中分布广泛、类型丰富。目前的一些句法分析器(如 Stanford parser 和哈尔滨工业大学的语言技术平台 LTP-Cloud)对比较句的处理没有应用到汉语教学领域。本文的语料主要来源于对外汉语教材，基于 HNC 理论，通过比较规则来识别比较句，最后的语义分析可以直接应用于对外汉语教学领域。

2 比较句类别

比较句在中文里分布广泛，但并非所有表示比较意义的句子都是比较句，本文比较句的筛选排除了与比较范畴邻近的判断范畴和选择范畴。最后，我们选择 6 种形式化的比较句作为研究目标，它们是“比、不比+……+A”“有、没有、有没有+……+A”、“像、和、跟、同+……+一样、差不多(+A)”、“最、还、更+A”、“A+于”以及“不如”类。我们的语料主要来源于“汉语国际教育动态语料库”中 14 本对外汉语教材的课文语料和 HSK 考试试题语料，共 68946 句。经统计，大约有 3000 个比较句。表 1 是 6 种类型比较句的语义以及它们在本文语料中的分布情况。可以看出，比较句在对外汉语教材的语料中分布广泛。

本文定义的比较句包含或隐含 4 项元素：比较项(包括比较主体和比较客体)、比较标记、比较点和比较结果，分别用 I，M，P 和 R 表示。这 4 项元素是提取比较关系的关键，同时也是比较关系抽取的关键。比如在句子“大商场的东西比小商店的东西贵”中，比较关系如表 2 所示。

2.1 比较标记(M)

比较标记是比较句最重要的成分，是判断一个句子是否为比较句的最显性的标记，在我们的调查中，无标记比较句在真实语料中所占比重非常小，

表 1　比较句的类型、语义和分布

Table1　Distribution and semanteme of comparative sentences

类型	语义	分布/%
比、不比+……+A	表示事物间性状、程度的差别或高低。	17.5
有、没有、有没有+……+A	表示事物达到某一度量范围。	2.2
最、还、更+A	表示事物更胜一筹。	68.5
A+于	表示事物在某一方面超过或劣于另一事物。	0.8
像、和、跟、同+……+一样、差不多(+A)	表示某一事物相同或相近于另一事物。	9
“不如”类	表示某一事物比不上或不同于另一事物。	2

表 2　比较关系举例

Table 2　Example of the comparative relation

比较主体	比较客体	比较标记	比较点	比较结果
大商场	小商店	比	东西(商品)	贵

本文的研究也排除无标记比较句。语言学家也多是通过比较标记的不同给比较句划分不同句式。同样, 本文的研究也是以比较标记的类别为纲。根据 HNC 理论, 把比较标记分为 5 类: L0(引导句子主块)、L1(引导句子辅块)、EG(句子核心动词)QE(核心动词前修饰性成分)和 HV(核心动词的后缀性成分), 表 3 是主要的比较标记。

2.2　比较结果(R)

比较结果是比较句必不可少的成分, 比较结果的不同也是造成比较句形式多样、结构复杂的原因。比较结果的提取是比较关系提取的重点, 对情感分析意义重大。表 4 是比较结果的语法表现形式。

2.3　比较句分类

综合考察比较标记和比较结果, 以比较标记为纲, 把比较句分为 6 类, 考察每一类的比较结果, 根据比较结果的不同分布, 可以划分为 33 个子类, 具体分布如表 5 所示。由于否定形式的比较句“不比”和“没有”有自己特殊的比较结果分布规律, 故把这两类单独拿出来考察其比较结果。

表 3　比较标记

Table 3　Comparative marks

HNC	比较标记词	例句
L0	比、不比、有、没有、有没有	你当然比我知道的多。
L1	像、和、跟、同……	你跟我妈妈一样高。
QE	更、还、最……	这位服务员的声音更大。
HV	于、过……	在衣食住行四事之中, “食”的程度高于其余一切。
EG	不如……	这个饭馆儿的菜是不错, 但是不如我们杭州的饭馆儿。

表 4　比较结果的类型

Table 4　Types of comparative result

类型	解释	例句
A	形容词谓语	你当然比我知道的多。
A + 补语	形容词谓语后有 HV 或数量补语	他们的水平比我们高多了。
VP	谓语为表心理状态的动词	我比你多了解一点儿情况。
VP+得+NP+A	谓语是一般动词后加形容词情态补语	她英语说得比普通话好。
增加类 V+ 数量补语	谓语是表示增加或减少的动词	每亩产量比去年增加了 50 斤。
先、后……+V+数量补语	谓语前有先、后等性质形容词	我比你早一点来到中国。

表 5 比较句类别 W

Table 5 Categories of comparative sentence

R	A	A +补语	VP	VP+得+NP+A	增加类 V+ 数量补语	先、后……+ V+ 数量补语
比	√	√	√	√	√	√
不比	√	√	√	√	no	no
有	√	no	√	√	no	no
没有	√	√	√	√	√	√
更、还、最	√	√	√	√	no	no
于、过	√	no	no	no	no	no
像、和、跟、同	√	√	√	√	no	no
不如	√	√	√	√	√	no

说明: √表示有此类型子句, no 表示没有此类型子句。

3 研究方法

我们的研究方法综合了比较句语料库、比较词词典和识别规则, 采用词驱动和语义规则相结合的策略进行比较句识别和分析。

3.1 流程

由于比较句的分布不是均匀的, 我们需要对语料库进行预处理, 包括人工过滤比较句、分词、词性标注以及建立知识库。然后, 根据比较句的不同性质建立不同的识别模型。在各个模型中, 都有识别比较句的规则, 把判定句子是否为比较句的规则放在比较句识别规则库, 如果一个句子成功匹配此规则库中的规则, 则判定这个句子是比较句, 否则就是非比较句。在此过程中, 比较句的句法分析也通过匹配规则完成。

3.2 预处理

3.2.1 比较词知识库

根据比较标记和比较结果, 我们建立比较词知识库, 知识库中的词语属性包含识别比较句的激活信息, 这些语义信息来自于 HNC 理论。比较词知识库来源于一个 30 万词 HNC 通用知识库。表 6 是抽取的比较标记词和比较结果词知识库。其中, CC 是概念类别, GCC 是广义概念类别, LV 是激活信息, GBK_NUM 是语义块的数量, GXGY 是句类属性, FRAME_VALUE 是语义格标。

3.2.2 消歧策略

很多比较词的概念类别不止一个, 但在特定的语境里, 一个词语只能有一种概念类别。由于比较词的概念类别和句子是否为比较句有直接关系, 所以对比较词汇的消歧处理就是至关重要的一步。例如, “比”字, 可以是一个动词, 也可以是介词, 做介词的时候又有 l1 和 l0 的区别, 只有当概念类别是 l0 时, 句子才是比较句。在处理每一类比较句之前, 我们都设置一些消歧规则, 使比较词能够识别正确的属性。

3.3 识别模型

根据以上 6 种类型比较句句法和语义特征的不同, 我们建立识别比较句的 4 类处理模型: L0 模型,

表 6 比较标记词和比较结果词知识库

Table 6 Wordlist of comparative mark and result words

词汇	属性
有	$ CC[v] GCC[V] LV[V] GBK_NUM[3] GXGY[GY] EPER[Y] SC[R] $ $ CC[hv] GCC[L] LV[HV] $ $ CC[l02] GCC[L] LV[L0] $
跟	$ CC[l02] GCC[L] LV[L0] $ $ CC[l14] GCC[L] LV[L1] $ $ CC[l4] GCC[L] LV[L4] $
相同	$ CC[u] CC[ug] GCC[U] LV[U] SC[jD000] FRAME_VALUE[0] $
增加	$ CC[v] GCC[V] LV[V] GBK_NUM[2] GBK_NUM[3] GBK_NUM[4] 4KUAI_HV[2] GXGY[GY] $

ABK 模型, EG 模型和 QE&HV 模型。

3.3.1 L0 模型

L0 模型是比较句处理中最典型的模型, 这一类模型主要处理的比较句有“比、不比+……+A”和“有、没有、有没有+……+A”以及“像、和、跟、同+……一样、差不多+A”3 种主要的比较句。在这 3 种结构里, “比、有、没有、有没有、像、和、跟、同”等词语的概念类别是逻辑概念 l0。

例句 1 中秋节有春节那么热闹吗?

第 1 步: 消歧规则。

“有”在比较句里只能是 l0 属性的词语, 所以“有”的其他属性需要排除, 规则 1 就是针对此种语言现象的排除规则。

规则 1 (0){CHN[有, 没有, 有没有]&LC_CC[v]}+(f){(1)CHN[这么, 那么]+(2)LC_CC[u]}=> !LC_SELECT(0, LC_CC, v)&!LC_SELECT(0, LC_CC, jlu)&!LC_SELECT(0, LC_CC, hv)$。

在这句话中, 比较标记“有”是规则的激活点, 用节点(0)表示, CHN 表示中文词形。LC_CC 表示概念类别, !LC_SELECT()表示排除某项属性。规则 1 意思是: 如果词条“有”、“没有”或“有没有”后边能够找到词条“这么”或“那么”并且“这么”” “那么”后边紧邻一个形容词, 此时的“有”” “没有”” “有没有”就不会选择 v, jlu 和 hv 属性, 最后只保留 l0 属性。

第 2 步: 选择谓语。

HNC 理论认为, 一个句子只有一个核心谓语, 因此, 正确选择谓语是整个句法分析系统的核心步骤。这句话是形容词谓语句, 需要把形容词识别为谓语核心。

规则 2 (b){(–2)CHN[有, 没有, 有没有]}+(–1) CHN[这么, 那么]+ (0){LC_CC[u]&END%}=>LC_TREE(E, 0, 0)&PUT(fp, LC_E_SCORE, E_U)$。

规则 2 表示, 如果一个形容词在句尾, 它前边是“这么”或“那么”, 再往前能找到“有”” “没有”或“有没有”, 这时, 我们给这个形容词生成 E, 即句子的谓语, 并且赋权值 LC_E_SCORE 为 E_U, E_U 是做谓语的最高的权值之一, 保证“热闹”可选为谓语。

第 3 步: 生成 L0。

因为“有”是 l0 概念, 并且是比较标记, “有”只有在句法树上挂为 L0 才能有效识别为比较句。

规则 3 (0){CHN[有, 没有, 有没有]&LC_CC[l02, l0]}+(f){(1)END%&LC_CHK[E]&LC_E_SCORE[E_U]}=>LC_TREE(L0, 0, 0)&PUT(fp, LEVEL, 1)$。

在这句话中, “有”会识别为 L0 并挂为 LEVEL 为 1 的树, 说明这是句子层级的比较。如果 level 不是 1, 则可能是比较语义块。

第 4 步: L0 与谓语相匹配。

L0 和谓语都识别出来后, 最后的关键一步是把 L0 与 E 匹配起来, 这时 E 会被 L0 赋上 E_FORMAT 的权值,才会最后被选为比较句的谓语。

规则 4 (b){(–1)CHN[有,没有,有没有]&LC_CHK[L0]&!LEVEL[2]}+(0){LC_CHK[E]&LC_E_SCORE[E_U]&END%}=>PUT(–1, LEVEL, 1)+PUT(0, LC_E_SCORE, E_FORMAT)$。

通过以上 4 步, 整个句子被分析为: 春节–GBK1, 有–L0, 中秋节–GBK2, 那么热闹吗–EG。GBK1 是比较主体, L0 是比较标记, GBK2 是比较客体, EG 是比较结果。

3.3.2 ABK 模型

利用 L0 模型可以处理“像、和、跟、同+……一样、差不多+A”结构, 但如果此结构中没有 A, 句子的核心谓语就会由“一样、差不多”这类词语充当, 整个句子的句法结构也会因此发生变化, 这时需要用 ABK 模型来处理。ABK 表示句子的辅助语义块, ABK 模型与 L0 模型类似, 不同之处是 ABK 模型的比较标记是逻辑概念 L1, 并且在句法树上, L1 不会对核心动词赋值, 核心动词可以直接生成 E 做谓语。因此, 与 L0 模型相比, ABK 模型少了最后匹配赋值的步骤, 而最为关键的一步变成对 ABK 辅块的识别。

例句 2 我们跟上一代人的想法不一样。

类似于例句 1, 首先对“跟”进行消歧处理, 使之选择 L1 属性; 然后进入辅块识别环节, 根据规则 4: (b){(–1)CHN[像, 与, 和, 同, 跟]}+(0)LC_CHK[L1H]&CHN[相比, 相比较, 一样]=>LC_TREE(L1, –1, –1)+LC_TREE[ABK, –1, 0)$, 使“跟……想法”生成 ABK; 最后, 识别并选择“不一样”为句子谓语。在规则 4 中, LC_TREE[ABK, –1, 0)表示从(–1)节点到(0)节点生成辅块 ABK, 其中“[ ]”表示实边界,“( )”表示虚边

界。例句 2 句的句法分析结果为: 我们GBK1比较主体, 跟上一代的想法ABK2比较客体+比较点, 不一样EG比较结果。

3.3.3 EG 模型

L0 模型和 ABK 模型处理比较标记为介词的情况, 有时候动词也可以作比较标记, 如“不如”类比较句。相较于以上两类模型, 此类型比较句的识别较为简单。在句法结构层面, 重点在于识别并选择“不如”类动词作谓语。由于作比较标记的动词是有限的, 可以穷尽。除了“不如”, 能够作比较标记的动词还有“不及”、“等于”、“赶不上”” “赶得上”” “超越”” “犹如”” “等于”等。只要能够正确识别句法结构就可以判断为比较句。但此类型比较句的比较结果却形式多样, 涉及到比较语义块和比较句的蜕化形式, 在复杂长句中处理效果还不理想。

3.3.4 QE&HV 模型

比较句结构“还、更、最+A”和“A+于、过”中的比较标记“还、更、最”是修饰动词的前加成分 QE, “于、过”是动词后缀 HV。这两种类型比较句识别的特殊之处在于, 它们是依附于核心动词 E 的成分, 属于谓语成分的复杂构成, 因此没有单独的处理步骤, 是在生成 EG 的步骤内完成的。同样, 句法识别的关键是 EG 的识别, 在识别阶段, 只要句子符合“还、更、最+A”和“A+于、过”结构就判断其为比较句。

4 实验及评估

4.1 实验语料

实验的测试语料来自对外汉语教学领域, 包括两部分: 一是 14 本经典对外汉语教材; 二是到目前为止的 HSK 考试的所有语料。共近 7 万句, 其中比较句约有 3000 句。

实验采用准确率、召回率和 F 值作为评测指标:

$$\text{准确率} = \frac{\text{识别正确的数量}}{\text{识别的数量}},$$

$$\text{召回率} = \frac{\text{识别正确的数量}}{\text{应该识别正确的数量}},$$

$$\text{F值} = \frac{\text{准确率}\cdot\text{召回率}\cdot 2}{\text{准确率}+\text{召回率}}$$

4.2 实验结果

本文的 HNC-system 系统是基于 HNC 语义理解理论建立的一套规则系统, 此系统的主要功能是句法语义分析, 分析结果可以应用于不同的自然语言处理需求。比较句识别规则首先根据 HNC-system 调用步骤分布于不同的模块, 对应于 4.3 的处理模型。这些规则既是系统句法语义分析的依据, 又是系统识别比较句的激活点。

根据 3.3 节比较句类型, 给出系统对比较句的句法分析结果, 见表 8, 我们参照哈尔滨工业大学

表 8 句法分析结果

Table 8 Result of parser

比较句类型	HNC-system		LTP-Cloud	
	准确率/%	召回率/%	准确率/%	召回率/%
比、不比+……+A	92.31	82.07	91.11	78.85
有、没有、有没有+……+A	92.98	88.33	65.85	46.56
像、和、跟、同+……+一样、差不多(+A)	98.11	88.14	75.00	55.10
最、还、更+A	96.56	89.58	98.35	91.67
A+于	96.88	81.49	97.35	87.35
“不如”类	96.72	82.87	85.71	73.47

表 9 识别结果

Table 9 Result of identification

	准确率/%	召回率/%	F-测度/%
SS+DR+SVM[9]	85.4	88.2	86.8
Keywords, Entity, SCR[10]	96.55	88.63	92.43
HNC-system	95.59	85.41	90.22

的语言技术平台 LTP-Cloud 进行比较。表 9 为张辰[9]将句法结构模板(SS)、依存关系相似度计算(DR)和 SVM 三者结合的识别结果, 黄高辉等[10]综合了关键词(Keywords)、实体(Entity)、CSR 的识别结果以及本文 HNC-system 对比较句的识别结果。

从表 8 可以看出, HNC-system 对各种类型的比较句分析的准确率都可以达到 90%以上, 召回率在80%以上。虽然“最、还、更+A”和“A+于”两类的效果稍逊于 LTP-Cloud, 但整体识别效果优于 LTP-Cloud, 并且各类型比较句识别效果稳定。其中“有、没有、有没有+……+A”的识别效果优于 LTP-Cloud, “像、和、跟、同+……+一样、差不多(+A)”和“不如”类的识别效果也得到显著提高。表 9 是 HNC-system 对比较句的识别结果和两个基于统计的方式对比较句识别结果的对比, 我们发现 HNC-system 可以达到甚至超过统计分析的效果, 这为未来统计与规则相结合来研究比较句的识别奠定了基础。HNC-system 对比较句的识别结果是在对比较句句法语义分析的基础上进行的, 为后续的比较关系抽取提供了基础。另外, 本文虽然按类型分别测试识别效果, 但交叉类型小类识别错误不影响整体识别效果, “更、还、最+A”与“比、不比+……+A”结构的交叉, 只要符合某一规则就会被识别为比较句。例如, 在“有的工作妇女比男人做得更好”中, 虽然“更好”收词影响“最、还、更+A”类识别结果, 但通过“比”字规则仍可以识别为比较句。

4.3 错误分析

通过实验发现, 关键词的激活和语料规模会影响比较句的识别结果。

1) 分词结果和知识库收词影响识别效果。利用规则的方式识别比较句, 除了需要精准分析句法和语义信息外, 还在很大程度上依赖于关键词的激活。分词错误会导致正确的规则无法被系统调用。另外, 目前的 HNC-system 采用的是 HNC 通用知识库, 有些关键词语被知识库收词后导致比较句无法被识别出来, 尤其是“最、还、更+A”和“A+于”比较句。例如“更好”“最好”等被收词后虽然不会影响句法分析的结果, 但会影响比较句的识别结果。应该在后续工作中做好 HNC 通用知识库的净化工作。

2) 语料规模影响识别效果。目前语料规模较小, 比较规则比较粗略, 有待进一步细化和规范化。本文的规则是根据 14 本教材中比较语法项目的例句进行总结的, 测试用的语料是从教材的课文里抽取的, 所以测试语料与例句会有小部分重合, 但这对实验数据的影响不大。如果扩大语料规模, 可以忽略这一影响。另外, 例句的比较句是比较典型的, 容易总结规律书写规则, 但在实际语料中, 比较句的句法形式复杂多变, 规则难以覆盖所有的形式, 甚至某些规则会制约比较句的正确识别。

5 总结和展望

本文以 HNC 理论为指导, 以面向汉语国际教育的教材语料为研究目标, 分析了比较句的类型及特征, 提出了一种基于规则的比较句识别方法。实验结果表明, 这种方法能有效识别比较句并分析比较句的句法成分。句法结构识别与句法分析是自然语言处理的基础性工作, 比较句的识别可以更好地服务于汉语国际教育, 在以后的工作中, 可以通过比较句类型的分布计算、比较标记词与比较结果的搭配信息计算等定量研究, 服务于汉语国际教育动态语料库语言资源建设以及检索系统的研发, 并为优化比较句的教学设计提供数据参考。

另外, 由于规则的方式开销较大, 有些类型的比较句通过规则可以大大提高识别的准确率, 但某些类型的比较句用统计的方式可以更方便快捷的识别正确, 因此, 未来的工作方向是规则与统计的结合。通过实验分析可以发现, L0 模型和 ABK 模型用规则的方式可以达到很好的处理效果, 而 EG 模型和 QE&HV 模型用规则的方式可以较好完成句法分析的任务, 但在比较句识别阶段, 可以考虑结合统计的 SCR 和模板匹配的方式, 在保证识别准确率的基础上, 提高效率。同时, 我们要扩展语料规模, 细化比较规则。在比较句识别和分析正确的基础上, 抽取比较关系是我们下一步的重点工作。

参考文献

[1] 丁崇明. 现代汉语语法教程. 北京: 北京大学出版社, 2009: 81–95

[2] 刘月华, 潘文娱, 故韡, 等. 实用现代汉语语法. 北京: 商务印书馆, 2001: 833–851

[3] 尚平. 比较句系统研究综述. 语言文字应用, 2006(S2): 77–80

[4] 车竞. 现代汉语比较句论略. 湖北师范学院学报: 哲学社会科学版, 2005, 25(3): 60–65

[5] 陈珺, 周小兵. 比较语法项目的选取和排序. 语言教学与研究, 2005(2): 22–33

[6] Jindal N, Liu B. Identifying comparative sentences in text documents // Proceedings of SIGIR. NewYork: ACM, 2006: 244–251

[7] Jindal N, Liu B. Mining comparative sentences and relations // Proceeding of AAAI. Palo Alto, 2006: 1331–1336

[8] 黄小江, 万小军, 杨建武, 等. 汉语比较句识别研究. 中文信息学报, 2008, 22(5): 30–38

[9] 张辰, 冯冲, 刘全超, 等. 基于多特征融合的中文比较句识别算法. 中文信息学报, 2013, 27(6): 110–116

[10] 黄高辉, 姚天昉, 刘全升. 基于 CRF 算法的汉语比较句识别和关系抽取. 计算机应用研究, 2010, 27(6): 2061–2064