基于多特征融合的中文比较句识别算法 *

张辰，冯冲，刘全超，师超，黄河燕，周海云
（北京理工大学，北京，100081）

摘要：观点承载着文本的重要信息，而比较句是观点评论中一种常见的句式现象。针对中文比较句识别问题，文章提出了一种基于规则与统计相结合的方法并进行实验。该方法先对语料及其分词结果进行规范化处理，再通过基于比较特征词词典与句法结构模板、依存关系相结合的方法进行泛提取。然后设计一种 CSR 规则提取算法，并利用 CRF 挖掘实体对象信息及语义角色信息。最后利用 SVM 分类器，选取不同特征维数，找到使性能达到最优的特征形式完成精提取。

关键词：比较句；规则；CRF；SVM
中图分类号：TP391　　**文献标识码：**A

Chinese Comparative Sentence Identification Based On Multi-feature Fusion

Zhang Chen, Feng Chong, Liu Quan-chao, Shi Chao, Huang He-yan, Zhou Hai-yun
(Beijing Institute of Technology, Beijing 100081, China)

Abstract: Opinions always carry important information of texts. Comparative sentence is a common way to express opinion. This paper described how to recognize comparative sentences from Chinese text documents by applying rule-based methods and statistical methods as well as analyze the performance of these methods. This method firstly normalized the corpus and its segmentation results, and then got the broad extraction results by using a lexicon-based method, sentence structure and dependent relationship analysis. Then a kind of CSR rule extraction algorithm was designed to extract the dependency relationship. The paper also used a CRF algorithm to identify entities and semantic roles. Finally, by using SVM classifier and choosing different feature dimensions the paper found the most optimum and effective features combination to finish the accurate extraction.

Key words: comparative sentence；rule；CRF；SVM

1　引言

几乎每天人们都被形形色色的选择所包围。为了做出更好的抉择，我们往往会选择拿我们感兴趣的物品作比较。在如今这个大数据时代，我们会从中得到海量的信息，这比传统的问卷调查式方法要好很多。然而与此同时我们却又为之困扰，同时处理这么大量的信息会是一件费时费力的事情。因此，我们需要一种比较观点挖掘系统来帮助我们自动从海量数据中得到两者（或更多事物）间的比较信息，这是一项有实用意义和学术意义的研究课题。

英文方面，文献[1]讨论了如何从英文文本中识别比较句，采用支持向量机（SVM）和

* **基金项目：**国家重点基础研究发展计划（973 计划）资助项目（2013CB329605，2013CB329303）；国家自然科学基金项目（61201351）；国家自然科学基金重点项目（61132009）

作者简介：张辰（1988—），男，硕士生，主要研究方向为自然语言处理、网络信息抽取等；冯冲（1977—），男，博士，副研究员，硕士生导师，主要研究方向为社会计算、机器翻译。

CSR（class sequential rules）算法识别比较句，达到了 84%的准确率以及 83%的召回率。文献[2]在文献[1]的基础上，又利用 LSR（label sequential rules）算法对比较元素进行抽取，取得了不错的效果。文献[3][4]利用 Web 搜索获取相关信息，借而比较两个对象，获得他们之间的关系。文献[5]依靠建立的规则从论坛抽取相应产品名称和属性从而进行比较。文献[6]基于模式识别的方法，提出通过特征抽取模板（IEP）将差比问题句的识别及其比较对象的抽取这两个任务合二为一同时进行并达到预期效果。

在中文领域，北京大学的黄小江等人[7]提出中文比较句的识别问题，在 Nitin 等人研究基础上，利用特征词、CSR 等作为 SVM 分类器特征，将中文比较句识别视为二分类问题。此后，黄高辉等[8]在文献[7]研究基础上，以 SVM 为分类器，以特征词和 CSR 序列规则为特征，同时利用 CRF 算法抽取实体对象，并增加以实体对象的信息（主要是对象的位置和数量）作为特征，对比较句进行识别，最终取得了 96%的准确率和 88%的召回率。文献[9]通过 HNC (Hierarchical Network of Concepts)理论实现了中文比较句的识别及其翻译的过程。

总的来说，比较句与比较关系识别的研究尚不系统和成熟，目前还处于起步阶段。而中文的句式更加灵活多样，因而中文比较句的研究相对更加困难。这些技术相对成熟，本论文的研究也借鉴了其中的一些思路和方法。目前识别的思路大多是模板匹配或者将该问题归类为机器学习问题，利用特征提取并构造分类器将句子划分为比较句与非比较句两类。同比较句与比较关系识别相关的处理技术有文本分类、实体抽取、情感分析等。本文通过利用规则泛抽取和分类精抽取两个步骤，并选取多种特征训练 SVM 分类器来进行自动识别中文比较句，最终取得了较好的效果。

2 汉语比较句概述

一般说来，比较句是含有比较和对比含义的陈述句，在语义上要求形成两个或多个对象的比较。按照车竞[10]的定义，现代汉语比较句是指谓语中含有比较词语或比较格式的句子。汉语比较句的句子结构通常包括四个基本比较元素，即比较主体、比较基准、比较点和比较结果。按照文献[8]的做法，本文将比较主体和比较基准称为比较实体对象，比较点称为比较属性。同时，此四元组也构成了比较关系，比如“诺基亚 N8 的屏幕不如 iphone 的好”，这句很明显是比较句，并可以表示为四元组<诺基亚 N8，iphone，屏幕，好>。在实际应用中，这四个比较元素有时并不会同时出现。

比较句的类型多种多样，语义语用复杂多变，目前在学术界关于比较句的定义和分类标准尚无定论。本文采用 COAE2013 评测标准中的划分方法，如下所示：

（1）差比（分级）。两者之间有顺序上的差异，句子中说明某一事物比另一事物好。

例如：<Doc1>诺基亚 N8 的屏幕不如 iphone 的好

（2）差比（不同）。只是说明两个事物有差异，但没有高低、优劣之分。

例如：<Doc2>途安和毕加索的风格特点、细致程度以及技术含量都存在差异。

（3）平比（相等或类似）。句子中两件事情具有相同的倾向或近似相等。

例如：<Doc3>诺基亚 N8 与 iphone 的通话质量差不多。

（4）极比(最高级)。多者之间的极值，在句子中说明一个事物是最好的或者最不好的。

例如：<Doc4>iphone4s 屏幕是目前所有手机中屏幕最好的。

（5）无比较词，但句子用来比较两个或者多个实体的特征，只是没有明确对他们分级。

例如：<Doc5>诺基亚 N8 的屏幕材质是 TFT 的，但是 iphone 屏幕的材质是 IPS 的。

此外还有一些比较句，由于人工标注争议大或按照商品本身时间顺序进行比较，故不在本次研究范围中。例如，比拟句、形如“越…越…”、“越来越…”的“递比句”等。

综上可知，只有在对比较句的定义、分类、句法结构等做全面科学的解释基础上我们才可以有效地提出自动识别的方法。同时结合比较句特点，利用比较句特征才是进一步提高识别准确率的良方。

3 汉语比较句识别

本文的汉语比较句识别方法处理流程如图 1 所示。

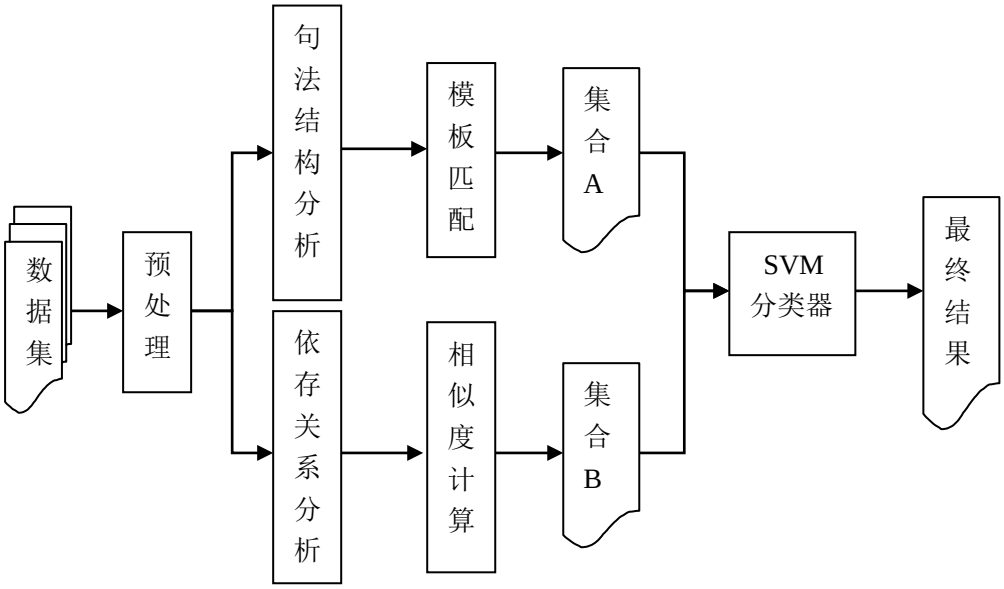

图 1 比较句识别方法流程

Fig.1 The process of comparative sentence recognition

该方法首先对语料进行规范化预处理，在一定程度上解决了语料不规范的问题。然后根据泛提取和精提取相结合的方法进行比较句识别。其中泛提取主要应用句法结构模板抽取以及依存关系相似度计算来分别识别显性/隐性比较句，得到的结果集 A+B 准确率低、召回率高；再经由精提取，利用 SVM 分类器对比较句进行抽取，得到的最终结果在不损失准确率的前提下召回率得到显著提升。

3.1 语料预处理

本文从 COAE 提供的电子、汽车两大领域各 1200 句训练集入手，通过对这些数据的分析，本文总结出如下特点：

- 语料数据以句为单位，维数稀疏、文本长度较短
- 比较句与非比较句数量比例严重不平衡，须进行平衡处理方能进行后续工作

• 训练语料中比较句对于比较关系四大基本元素并不完整，或很隐晦

• 语料领域性极强，且口语化严重

针对以上特点，预处理的具体步骤如下：

1) 使用中科院计算所开发的NLPIR2013[1]对语料分词和词性标注，并将分词结果与领域名词词典以及比较特征词词典中的词进行比对，校对标注结果；
2) 使用斯坦福大学的Stanford Parser[2]进行句法结构分析，尤其是以比较特征词为核心，对句中主语、谓语以及谓语同根子节点进行正确划分。
3) 使用哈工大的LTP[3]对语料进行依存关系分析及语义角色标注。
4) 针对语料数据不平衡问题，参考文献[11]提出的熵值平衡算法进行平衡处理，得到接近 1:1 的平衡语料。

经过这四步与处理流程，有效提高挖掘精度，为最后整个挖掘的成功奠定了基础。

3.2 句法结构模板抽取

如前文所述，目前关于比较句识别方面有基于规则（CSR、模板库）以及基于统计（SVM）的方法，但鲜有二者相结合的方法。但是不管采用何种方法都会势必造成一定数量的错判。本节通过研究语料及日常比较句语言特点，发现了一些经常被人们用于比较的表达方式，并通过验证试验最终归纳出覆盖度较高的句法结构模板。

汉语是一门高度灵活多变的语言。尽管大多数比较句会包含比较特征词，如“比”、“不如”、“一样”等；但也有一些句子不会包含这些词（通常为差比），例如“诺基亚 N8 的屏幕材质是 TFT 的，但是 iphone 屏幕的材质是 IPS 的。”，从表面上看它是一个转折句，但实际表达的确是比较的含义。

几乎所有比较句都有比较特征词，文献[12]列举了一些中文常用比较词以及比较结果词，如表 1 所示。

表 1 中文常用比较词及比较结果词表

Tab.1 The comparative words commonly used in Chinese

比较词	比较结果词
比，比起，相比，对比，较，较之，比较，不如，不及，比不上，亚于，逊色于，劣于，弱于相当于，等于，等价于，近似于，像，犹如，如同，堪比优于，强于，好于，高过，胜过，超过，超越，有别于区别于	差异，差距，差别，区别，不同，不一样，一样，媲美，雷同，相同，分庭抗礼，不相上下，旗鼓相当，差，弱，欠佳，欠缺，不足，劣势，缩水，稍逊一筹，好，强，飞跃，提高，提升，增加，增强，进步，领先，改善，改进，扩充，优势，更胜一筹，佼佼者，首屈一指，出类拔萃

使用这些比较特征词会大大提高我们最终结果的召回率，为我们之后的工作打下基础。与此同时我们还要兼顾那些没有出现比较特征词的句子，因此我们给出了如下的定义。

定义 3-1：以含有比较特征词，明确表达两者（或多者）之间对比的句子，称为显性比较句。

例如：诺基亚 N8 的屏幕不如 iphone 的好。

[1] http://www.nlpir.org/
[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] http://ir.hit.edu.cn/ltp/

定义 3-2：不含有比较特征词，但整体意图是为了比较两者（或多者）之间的特征的句子，称为隐性比较句。

例如：诺基亚 N8 的屏幕材质是 TFT 的，但是 iphone 屏幕的材质是 IPS 的。

本节主要对显性比较句进行研究。通过观察大量语料，本文将显性比较句的句法结构总结为如下三种模式，这三种模式都是以比较特征词作为匹配的起始点：

1）SS_1 = ... + VP (Keywords/Key Phrases) + ...VA/ADJP...

此模式含义为：句子中出现了比较特征词，且此特征词的父节点为 VP，其父子节点中存在表语形容词或形容词短语

2）SS_2 = ... + VP (Keywords/Key Phrases) + ...ADVP...

此模式含义为：句子中出现了比较特征词，且此特征词的父节点为 VP，其父子节点中存在副词短语

3）SS_3 = ... + NP (Keywords/Key Phrases) +...

此模式含义为：句子中出现了比较特征词，且此特征词的父节点为 NP

此外，为了保证比较句的识别准确率，提取比较句的词性、位置、语义等特征也将对识别效果的提升产生帮助。

3.3 依存关系相似度计算

如上文提及，我们将比较句分为显性比较句和隐性比较句并分别进行了定义。利用 3.2 中的句法结构模板可以识别出召回率较高的显性比较句，而隐性比较句由于其语义复杂性，我们希望透过依存关系来挖掘其中更多的有效信息。依存句法分析系统用于对汉语进行句法分析，将句子由一个线性序列转化为一棵结构化的依存分析树，通过依存弧反映句子中词汇之间的依存关系，弧的方向是由核心词指向依存词，弧上的标一记表示依存关系的类型[13]。对于隐性比较句“诺基亚 N8 的屏幕材质是 TFT 的，但是 iphone 屏幕的材质是 IPS 的。”进行依存关系分析，解析效果如下

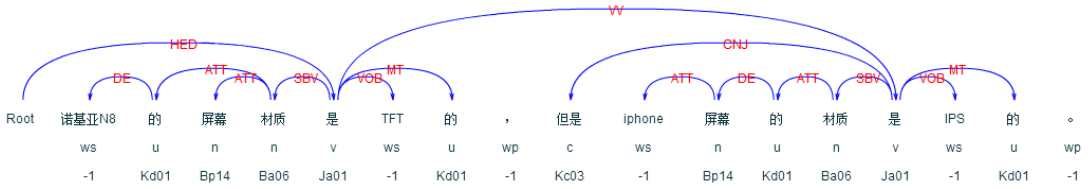

图 2 依存关系分析结果示意图

Fig.2 The diagram of dependency relationship analysis

通过观察这类隐性比较句，我们会发现其前后部分的依存分析结果会存在大量相似结构，而其连接处多半会以标点符号、转折词、并列词进行衔接。通过统计这类句子的依存关系，借鉴文献[14]中关于树相似度算法，计算前后两部分依存关系的相似度并设定阈值，判断其是否属于隐性比较句候选集。

此方法对于并列关系、转折关系类隐性比较句具有较好的提取效果，但须与其他方法

结合才能到最终高准确率和高召回率的抽取目标。

3.4 利用 SVM 分类器识别比较句

在 3.2 中我们提出了三种高覆盖度的句法结构模板来识别显性比较句，在 3.3 中我们通过依存关系相似度算法抽取出隐性比较句，他们均达到了极高的召回率，但也都存在着相同的缺憾，即准确率较低。为了达到高准确率和高召回率的比较句抽取结构，我们考虑将比较句识别看作一个二分类问题。在汉语比较句识别领域，文献[7][8][11]均采用 SVM 分类器进行比较句的判别并取得了不错的效果。因此本文也将使用 SVM 分类器，在前人基础上增加分类特征以达到提升分类效果的目的。

在分类特征选取上，本文认为比较句和非比较句分属两个截然不同的文本种类，此二者无论是在语义层面上，还是在语法结构上都会隐含着自身独有的特征。为此，我们共提出了以下四种特征作为 SVM 的候选特征向量：类别序列规则（CSR）、语义角色标注（SRL）、比较特征词以及统计词特征。

3.4.1 类别序列规则

由于在泛提取中使用的是句子层面的分析，而对于细粒度的词序列层面没有进行过多的挖掘，故在 SVM 特征的选择上，我们添加了另一模板类信息——类别序列规则。另一点考虑是虽然与句法结构同为模板特征，但句法结构在约束力上更为宽泛，更符合泛提取的要求，而 CSR 的轻便性以及约束力强，使其更适用于精提取的分类过程。此外在泛提取候选集上进行 CSR 规则挖掘分析也将大大提升这一特征抽取的准确性并提高整体效率。

序列模式挖掘(Sequential pattern mining, SPM)是数据挖掘中的重要任务之一。类别序列规则(Class Sequential Rule, CSR)[1]是序列模式的一种，把类别序列规则应用于比较句识别中与其它序列模式挖掘的思想一样，都是寻找满足用户定义好的最小支持度约束的模式，为后期的比较句识别提供特征输入。在文献[7][8]中均使用 CSR 规则作为分类特征，其中文献[7]提出以一个分句作为一个序列，这种做法在常规句中效果会比较好，但是对于口语化较严重以及不规范语法的句子将起到事倍功半的效果，故我们这里选用不同窗口长度分别进行实验。同时，由于样本稀疏问题导致出现 CSR 序列的最小频率以及频率排序中位数都是 2，故此处置信度阈值亦选取为 2。实验后发现以分句为窗口不仅效率低下而且会导致准确率下降，而选取窗口大小为 5 的序列刚好可在此两方面获得均衡，因此对于 CSR 挖掘我们限定其最大长度为 5 个元素。

单单使用这一特征不能很有效地区别比较句与非比较句，主要是由于词性并不能很完整地表达句子的完整语义，在文献[8]中作者给出了很好的实例进行验证，因而我们还需添加更多的语义信息来完善分类特征。

3.4.2 语义角色标注

语义角色标注（Semantic Role Labeling, SRL）是近几年来的研究热点，在 CoNLL2004，CoNLL2005 以及 CoNLL2008 的任务中均有出现。SRL 在自然语言处理中的主要任务包括识别句子中与动词或谓语相关的语义成分进行识别，并将它们分派到相应的具体的角色类别中，如:“施事者(Agent)”、“受事者(patient)”、“讲话者(Speaker)”等[15]。文献[16]将中文

比较观点句分为六个基本元素：观点持有者（Holder），比较实体 1（Entity1），比较词（Comparative predicates），比较实体 2（Entity2），比较特征（Attributes）以及情感倾向（Sentiments）。其中观点持有者是表达比较观点的人，比较实体是在比较句拿来比较的人或物或事。比较特征是实体与实体比较时的比较点，一个比较句中可能会出现多个比较特征且特征中蕴含属性。比较词是汉语中表达比较关系的词语，例如“不如”、“比”。情感倾向指对比较实体的区分态度。

通常来讲一个比较句中比较实体首先会是一个命名实体（例如人、商标、地点），比较属性是一个名词，情感倾向会是一个形容词。但是实际情况往往会比这复杂很多，尤其中文中有大量短语以及成语的出现，导致使用规则或模板进行匹配并不是一个很好的选择。本文将主要精力放在了识别前五个角色，采用的方法是使用有监督的机器学习过程。在比较句识别中我们无须对实体的具体边界进行识别，更重要的是获取其相对位置，因此本文选取主题识别中的主流标注算法 CRF，并利用已经标注好前五个角色的语料进行训练。

条件随机场（Conditional random fields，CRF）是 Lafferty 等人[17]于 2001 年，在最大熵模型和隐马尔科夫模型的基础上，提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。在 CRF 模型中，特征的选取至关重要。本文在这里选用的特征为词、词性、短语类型、比较词、与比较词的间距、领域词典、停用词词典。其中分词之后得到的词和词性将为 SRL 提供非常有益的帮助，比如形容词或者副词很有可能是情感倾向，而命名实体则很有可能是比较实体。短语类型是通过句法结构分析步骤得到，与词性类似比如 NP 结构也很有可能成为比较实体。比较词及其他词与比较词的间距将使比较实体的具体位置识别变成可能，分别出比较对象与比较基准。最后领域词典与停用词词典将大大提升最终标注的效果。

3.4.3 比较特征词以及统计词特征

比较特征词在 3.2 中已有所提及，这里不再赘述。所谓的统计词特征，是在类别平衡处理之后的数据集上进行，通过计算某一个词 t 在类内和类间的分布，就可以得到该词汇在给定的这个数据集合上的分布情况，选取类间信息熵小、类内信息熵大的词汇就可以作为该类别的统计特征。设 $p(c_i|t)$表示 t 出现在文本中时，文本属于类 c_i 的概率，则某一词汇在类 c_i 内的信息熵为 $E(t)=-\sum_{i=1}^{r} p(c_i \mid t)\log_2 p(c_i \mid t)$。其值越大，说明词 t 在类别 c_i 中出现越频繁，越能代表该类文本。最后计算出每个特征的信息增益值，通过设定阈值来过滤掉噪声特征，将剩下的大于指定阈值的特征作为最终统计词特征。

4 实验

4.1 数据集

目前对于汉语比较句的研究还很少见，没有较多公开的评测数据集。此次在 COAE2013 评测数据基础上，笔者又收集了一些评测数据并进行人工标注，数据来源于“中关村在线”等产品评论网站，包括新闻正文、用户评论及论坛数据三类，包括汽车领域和电子产品领

域。数据集样本情况如下表所示。

表 2 数据集样本情况

Tab.2 Sample data sets

数据	句子	汽车领域/句	电子领域/句	总计/句
训练	比较句	500	500	1000
	非比较句	3000	3000	6000
测试	比较句	400	400	800
	非比较句	1000	1000	2000

4.2 实验结果

4.2.1 利用句法结构模板识别比较句

由于句法结构模板主要针对显性比较句进行研究，因此我们首先要验证这三种模板在显性比较句中的覆盖率。本文实验数据集在不考虑领域情况下包含 1800 句比较句与 8000 句非比较句，在此覆盖率验证实验中，我们对这 600 句比较句进行了句法分析，分别统计出三种模板在显性比较句中出现的次数，并依次计算了其在显性比较句中的占用率及在比较句中的占用率。

通过实验发现，在 1800 句比较句中，显性比较句共有 1742 句，由此可见，在比较句的组成中，显性比较句基本占据了大多数；此外，本文总结的三种句法结构模板在显性比较句集中覆盖了 1742 句中的 1739 条，覆盖率高达 99.8%，这说明，此三种模板几近全面地概括了显性比较句中的特征。但在置信度检测方面由于这三种句法结构模板设置得比较宽泛，导致其准确率并不尽如人意，三者加起来只打到了 65.6%的准确率，这说明本文提出的句法结构模板需要与其他方法相配合才能达到高准确率和高召回率的抽取目标。

4.2.2 利用依存关系识别比较句

在本节实验中，我们使用哈工大的 LTP 依存关系分析模块对 1800 句比较句进行解析，并利用 3.3 中提到的相似度计算方法识别比较句。同样在 58 条隐性比较句下，此方法达到了非常高的召回率，达到了 100%；但是同样，和句法结构模板相类似，准确率方面只有 13.3%。这说明在单独识别隐性比较句方面依存关系能与其他方法进行配合达到高准确率与高召回率的抽取目标。

4.2.3 利用 CRF 进行语义角色标注

由于 CRF 训练过程中并未将比较句与非比较句进行区分，只是以角色标注上的缺失来代替，故所得实验结果精度并不是十分准确，尤其是在比较对象（即比较实体 2）的识别准确率只达到了 83.5%。但是本文主要任务着眼于比较句的识别而非比较句中的语义信息挖掘，在此步骤中获取到的实体信息对于我们后续步骤中的 SVM 分类已起到了足够多的效果。

4.2.4 利用 SVM 进行比较句识别

本文选取 SVM 分类器在文献[8]的研究基础上，将实体对象信息扩展为 4.2.3 中得到的 SRL 标注结果，并添加统计词特征。实验数据采用 4.1 中提及的数据集，实验结果如表 3

所示。其中 Keyword 表示以比较句特征词为特征，CSR 表示以 CSR 序列规则作为特征，SRL 表示以语义角色标注信息作为特征，WSF 表示以统计词特征作为特征。依次选取特征进行组合实验，结果如表 3 所示：

表 3 利用 SVM 识别比较句实验结果

Tab.3 The experiment results of identifying comparative sentences using SVM

Feature	Pre/%	Rec/%	F-measure/%
Keyword	69.5	72.3	70.9
Keyword+CSR	75.7	80.6	78.1
Keyword+SRL	76.8	79.1	77.9
Keyword+CSR+SRL	77.3	79.3	78.3
CSR	74.4	68.8	71.5
CSR+SRL	76.5	78.9	77.7
WSF	69.2	73.2	71.1
WSF+CSR	80.4	81.5	80.9
WSF+SRL	83.7	79.4	81.5
WSF+CSR+SRL	83.9	80.7	82.3
Keyword+WSF+CSR+SRL	**85.9**	**82.4**	**84.1**

由上表观察知，单单使用统计词特征对于句子的分析力度明显不够，而在加入 CSR 或 SRL 等句法、语义信息后将使得结果得到显著提升，在一定程度上说明了统计特征与序列特征具有互补性，也验证了比较句具有重要的语法特征。在单特征实验中 CSR 表现最佳，这表明比较句的主要语义信息都集中在以比较特征词为中心的窗口大小为 5 的范围内。最终组合实验结果表明采用四种特征相结合的 SVM 分类器能有效提高抽取精度，在准确率、召回率方面都有所提升。

4.2.5 泛提取与精提取组合实验

将句法结构模板（SS）、依存关系相似度计算（DR）、SVM 三者结合。先用句法结构模板进行显性比较句粗匹配、依存关系相似度计算进行隐性比较句粗匹配，两者作为泛提取的结果再用训练好的 SVM 分类器进行处理，最终完成精提取。分别对这三者进行组合实验，结果如下表所示：

表 4 泛提取与精提取组合实验结果

Tab.4 The experiment results of combining general extraction and accurate extraction

	Pre/%	Rec/%	F-measure/%
SVM	85.9	82.4	84.1
SS+SVM	84.8	87.6	86.2
DR+SVM	86.5	84.7	85.6
SS+DR+SVM	**85.4**	**88.2**	**86.8**

实验结果表明使用泛提取与精提取相结合的方法对抽取结果的提升是很明显的。当句法结构与 SVM 分类器相结合时，准确率有所下降，但召回率提高了；当依存关系与 SVM 分类器结合时，准确率召回率均有所提高；当三者进行结合时，比较句识别结果最佳，同时 F 值达到了 86.8%，虽然准确率略比单独使用 SVM 有所降低，但是召回率得到了大大地提高，最终结果得到明显改善。

5 总结与展望

本文针对句子级别的比较观点挖掘问题，尤其是汉语比较句识别进行了简要的介绍，提出了新的解决思路并进行验证。在 Jindal、黄小江和黄高辉等人的研究基础上，提出了一种通过模板提取（泛提取）与概率分类（精提取）相结合的比较句识别技术。在泛提取中利用特征词词典、句法结构提取显性比较句；接下来利用依存关系提取隐性比较句；最后利用多种特征构造 SVM 分类器进行结果的筛选。实验结果显示，该方法在 COAE2013 语料的抽取效果较好。然而，有些问题还有待更深入的研究，下一步工作中将重点探究如下问题：1）对现有的规则模板进行同义词扩展，改进 CRF 标注算法，尝试提出更具普遍意义的依存关系匹配算法。2）在实体识别中的指代消解等问题仍没有考虑，有待进一步从篇章级文本中获取信息。3）通过阅读其他文献，尝试使用不同分类算法对结果进行测试。另一方面，在汉语比较句语料库的建设上，收集一个更大规模的比较句集合也是势在必行。

参考文献

[1] Jindal N, Liu B. Identifying comparative sentences in text documents[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 244-251.

[2] Jindal N, Liu B. Mining comparative sentences and relations[C]//Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, 21(2): 1331.

[3] Sun J T, Wang X, Shen D, et al. CWS: a comparative web search system[C]//Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 467-476.

[4] Luo G, Tang C, Tian Y. Answering relationship queries on the web[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 561-570.

[5] Feldman R, Fresko M, Goldenberg J, et al. Extracting product comparisons from discussion boards[C]//Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007: 469-474.

[6] Li S, Lin C Y, Song Y I, et al. Comparable entity mining from comparative questions[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 650-658.

[7] 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, 22(5): 30-38.

[8] 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[J]. 计算机应用研究, 2010, 27(6).

[9] Zhang R, Jin Y. Identification and Transformation of Comparative Sentences in Patent Chinese-English Machine Translation[C]//Asian Language Processing (IALP), 2012 International Conference on. IEEE, 2012: 217-220.

[10] 车竞. 现代汉语比较句论略[J]. 湖北师范学院学报 (哲学社会科学版), 2005, 3.

[11] 李建军. 比较句与比较关系识别研究及其应用[D]. 重庆大学, 2011.

[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, 23(2): 102-107.

[13] 胡宝顺, 王大玲, 于戈, 等. 基于句法结构特征分析及分类技术的答案提取算法[J]. 计算机学报, 2008, 31(4): 662-676.

[14] 刘伟，严华梁，肖建国，等. 一种 Web 评论自动抽取方法[J]. Journal of Software, 2010, 21(12): 3220-3236.

[15] Wang S, Li H, Song X. Automatic Semantic Role Labeling for Chinese Comparative Sentences Based on Hybrid Patterns[C]//Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on. IEEE, 2010, 1: 378-382.

[16] Hou F, Li G H. Mining Chinese comparative sentences by semantic role labeling[C]//Machine Learning and Cybernetics, 2008 International Conference on. IEEE, 2008, 5: 2563-2568.

[17] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.