

基于条件随机域模型的比较要素抽取研究

王巍¹ 赵铁军¹ 辛国栋² 徐永东²

摘要 随着主观性评价文本数量的不断增长, 文本情感分析已经成为众多研究者关注的对象. 比较要素抽取是比较句情感分析的重要研究任务之一, 比较句的情感分析结果与比较要素相结合才更有意义. 为了提高比较要素抽取的性能, 本文提出在构建系统模型的过程中引入浅层句法信息、比较词候选信息和启发式位置信息等多种语言学相关特征, 并且在不断增加领域知识的情况下, 有效提高系统的准确率和 F1 值, 同时本文提出的方法可以有效处理含有多个比较关系的句子. 实验结果表明, 将本文提出的特征应用于条件随机域 (Conditional random fields, CRFs) 模型可以有效提高比较要素抽取的各项性能指标, 同时, 将本文的实验结果与 2012 年中文情感分析评测结果的最大值进行了比较, 各项指标均超过最大值, 进一步证明了本文方法的有效性.

关键词 情感分析, 比较要素抽取, 浅层句法特征, 比较词候选特征, 启发式位置特征

引用格式 王巍, 赵铁军, 辛国栋, 徐永东. 基于条件随机域模型的比较要素抽取研究. 自动化学报, 2015, 41(8): 1385–1393

DOI 10.16383/j.aas.2015.c140762

Extraction of Comparative Elements Using Conditional Random Fields

WANG Wei¹ ZHAO Tie-Jun¹ XIN Guo-Dong² XU Yong-Dong²

Abstract With the rapid growth of the number of evaluative texts on the Web, sentiment analysis has attracted the attention of researchers all over the world. Extraction of comparative elements is one of the important tasks for sentiment analysis of comparative sentences. It is more meaningful that results of sentiment analysis combine with comparative elements. To improve the performance of comparative elements extraction, this paper proposes to introduce shallow parsing features, comparative word candidates and heuristic position information to conditional random fields (CRFs) for building a system model. The proposed method is not only free from introducing domain knowledge but also can effectively deal with sentences containing a few comparative relationships. Experiment results show that the performance of system is improved when introducing proposed features to the CRFs model. Meanwhile, compared with the best results of the 2012 Chinese opinion analysis evaluation, the F1-scores of the proposed method are higher than the maximum value.

Key words Sentiment analysis, comparative element extraction, shallow parsing feature, comparative word candidate, heuristic position feature

Citation Wang Wei, Zhao Tie-Jun, Xin Guo-Dong, Xu Yong-Dong. Extraction of comparative elements using conditional random fields. *Acta Automatica Sinica*, 2015, 41(8): 1385–1393

近年来, 文本情感分析已经成为自然语言处理领域的热点研究问题. 情感分析又称为观点挖掘, 是

收稿日期 2014-11-02 录用日期 2015-04-11
Manuscript received November 2, 2014; accepted April 11, 2015
国家高技术研究发展计划 (863 计划) (2015AA015405), 国家自然科学基金 (61402134, 61173073, 61172099, 61272384), 国家国际科技合作专项 (2014DFA11350) 资助

Supported by National High Technology Research and Development Program of China (863 Program) (2015AA015405), National Natural Science Foundation of China (61402134, 61173073, 61172099, 61272384), and the Special Project of International Science and Technology Cooperation of China (2014DFA11350)

本文责任编辑 刘成林
Recommended by Associate Editor LIU Cheng-Lin

1. 哈尔滨工业大学计算机科学与技术学院机器智能与翻译研究室 哈尔滨 150001 2. 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001

1. Machine Intelligence and Translation Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001 2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

指通过对大量的评论性文本进行分析以获取观点持有者对于产品、服务、事件等实体的观点、态度和情感^[1-2]. 文本情感分析技术已经成功应用于决策支持、问答系统、自动文摘等自然语言处理相关任务中^[3-5]. 比较句的情感分析是文本情感分析的一项基本任务, 具有重要的研究价值.

比较句情感分析的处理对象是评价文本中的比较语句, 其任务就是对比较句中的观点性要素进行抽取与分析, 包括对每个比较关系的情感倾向性进行判别以及从中抽取与倾向性相关的各个要素, 这些要素包括比较实体、比较属性、比较词和评价词. 一个典型的比较句, 例如“宝马的发动机好于奔驰”表达了两个汽车品牌之间的一种比较关系, 该关系可以用一个五元组来表达: (宝马, 奔驰, 发动机, 好, 于), 其中, “宝马”和“奔驰”分别是比较主体和比较客体, “发动机”是比较属性, “好”是相对比较主

体的评价词,“于”是比较词.通过对宝马和奔驰在“发动机”这一共有属性上的比较,表达了作者对宝马车的褒义情感.

比较句情感分析的首要任务就是定位并抽取句子中的比较要素,然后针对抽取出的要素,判断作者对不同比较对象的情感倾向性.比较要素抽取与常规观点句的信息抽取不同,其抽取的评价对象是具有比较关系的实体及其共有属性,而并非作者直接评价的单一实体或属性,实体间的比较关系通常以比较词来体现,因此本文引入比较词候选特征和启发式位置特征以提高系统对比较实体的识别能力.比较要素抽取的另一个问题是如何完整地识别出短语级要素.例如,在产品评论中,产品名往往由品牌名与型号名共同构成,如果只提取其中某一项作为比较实体将影响信息抽取的完整性和准确性.为此,我们提出引入浅层句法特征以增强系统识别短语级要素的能力.另外,比较要素抽取还需要解决如何区分不同类型的要素的问题.例如,通常情况下,比较主体、比较客体和比较属性三类要素具有相似的词性(名词词性)难以区分,但它们与比较词之间的相对位置关系具有一定的指示作用.

综上所述,为构建一个通用的比较要素抽取系统,本文引入了浅层句法特征、比较词候选特征和启发式位置特征等语言学相关特征,在不增加领域知识的情况下,有效提高了比较要素抽取的各项性能指标,表明了本文所提方法的有效性.

本文后续内容安排如下:第1节介绍了比较要素抽取的研究现状;第2节介绍了基于条件随机域模型的比较要素抽取过程;第3节描述了系统中引入的各项特征;第4节详述了实验结果及分析;最后是结论及下一步工作展望.

1 比较要素抽取的相关研究

本文的研究任务与 Jindal 等^[6]和 Yang 等^[7]的研究工作类似,其目的都是获取与比较论述相关联的各个要素. Jindal 等首先在距离各个比较要素窗口半径为 k 的范围内挖掘词性标记序列规则(Label sequential rule, LSR) (如: $\langle \{*, NN\} \{VBZ\} \rangle \rightarrow \langle \{ \$entity1, NN \}, \{ VBZ \} \rangle$), 进而按照置信度由高到低的顺序使用这些规则标记各类比较要素. 而 Yang 等首先基于词性信息标记出比较要素候选, 继而在距离候选要素窗口半径为 k 的范围内构造词性序列模式, 并以这些序列模式为特征建立要素识别模型. 可以看出, 他们的工作是基于上下文词性信息来获取具有特定词性的比较要素. 然而, 他们的方法基于以下假设: 一个比较句中仅包含一组比较关系, 事实上很多时候该条件是不成立的. 因此, 这种方法具有较大的局限性.

另外, 一部分研究者通过建立领域词典来识别比较要素. Xu 等^[8]通过人工收集手机领域常用的产品名称和属性名称构建了相应的领域词典, 依据这些领域词典来获取相应的比较实体和比较属性, 对于评价词, 他们通过常用的情感词词典进行获取. Feldman 等^[9]手工构建了跑鞋和汽车领域的产品品牌词典和属性词典, 对于产品型号, 他们采用正则表达式进行获取. 但是很明显, 他们的方法领域依赖性太强, 不具有领域的可移植性.

为了更准确地获取各类比较要素, 一些研究者将比较要素抽取任务看作序列标注任务. Xing 等^[10]为了抽取标准技术比较句中的技术指标, 选取关键词、名词短语以及位置信息作为特征建立比较要素抽取模型. 黄高辉等^[11]首先识别比较句中的实体对象, 继而将抽取的实体对象作为特征, 结合定义的规则, 建立比较主体和比较客体的抽取模型, 对于比较属性和评价词, 采用简单的词典匹配的方法进行识别. 国内的宋锐等^[12]选取了词、词性、启发式规则和领域知识等作为特征, 通过提取各类特征训练出比较要素抽取模型, 并将该模型用于比较要素的抽取. 然而很显然, 他们的方法仍然需要依赖领域知识, 领域移植性较差.

此外, 一些学者通过 Bootstrapping 的方法来识别比较要素. Li 等^[13]针对比较问句中比较实体的抽取问题, 采用 Bootstrapping 方法迭代地抽取比较实体对 $\langle entity1, entity2 \rangle$, 从获取的比较实体对产生新的抽取模式, 进而将新的抽取模式用于获取新的比较实体对, 直到没有新的模式产生. Ding 等^[14]从若干种子实体出发, 迭代地抽取更多的实体, 每次迭代过程均使用了序列模式挖掘算法, 将获取的序列模式用于新实体的发现. 很明显, 他们的工作是弱监督的, 不需要对大量的语料进行标注.

在比较要素抽取的工作中, 还有一部分学者将要素抽取任务转换成语义角色标注任务. Wang 等^[15]根据比较句的句法结构定义了三种比较模式 (如: $\langle entity \rangle$, $\langle comparative \ word \rangle$, $\langle entity \rangle$, $\langle sentiment \ word \rangle$) 来描述各个比较要素之间的关系, 并将这些模式泛化为 6 条模式来标注比较句的语义角色. 李建军^[16]利用语义角色标注器和 Stanford 句法分析器构建了语义角色句法分析树, 并通过计算两棵子树间的匹配相似度来获取比较要素. 然而, 以上的工作虽然可以获取比较关系中的比较实体等要素, 但对于含有多个关系的比较句只能抽取其中的一个比较关系.

2 基于条件随机域模型的比较要素抽取

2.1 条件随机域模型

条件随机域模型 (Conditional random fields,

CRFs) 是由 Lafferty 等^[17] 于 2001 年提出的一种判别式模型, 常用于标注或切分序列化数据. CRFs 模型结合了生成式模型 (Generative model) 和序列分类模型 (Classification model) 的优点, 是一种无向图模型. 该模型的优点主要表现在: 在给定观测序列的情况下, 将整个标记序列的概率分布表达为一个统一的指数模型; CRFs 模型没有生成式模型那样严格的独立性假设, 便于使用复杂的、有重叠性和非独立的特征进行训练和推理, 使得该模型在融合多种特征方面存在优势; CRFs 模型解决了其他判别式马尔科夫模型难以避免的标注偏置问题 (Label bias problem), 通过对所有状态序列进行全局归一化来获得全局最优的标注, 因此我们使用条件随机域模型对比较要素进行标注.

条件随机域模型是在给定一组输入随机变量的条件下, 一组输出随机变量的条件概率分布模型, 其中线性链 (Linear chain) 条件随机域已经成功应用于信息抽取、句法分析等多个自然语言处理领域. 由于比较要素抽取任务可以看作序列标注问题, 因此, 本文采用线性链条件随机域对比较要素进行标注.

设随机变量 $X = (X_1, X_2, \dots, X_n)$ 表示需要标注的观测序列, 随机变量 $Y = (Y_1, Y_2, \dots, Y_n)$ 表示观测序列对应的标记序列 (也称为状态序列), 在给定输入序列 x 的条件下, 标记序列 y 的链式条件概率分布如式 (1) 所示:

$$P(y|x) = \frac{1}{Z(x)} \times \exp \left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right) \tag{1}$$

式 (1) 是线性链条件随机域模型的基本形式, 表示给定输入序列 x 对输出序列 y 预测的条件概率, 式中, t_k 和 s_l 是特征函数, λ_k 和 μ_l 是特征函数对应的权值, $Z(x)$ 是规范化因子. t_k 是定义在边上的特征函数, 称为转移特征, 依赖于当前和前一个位置, s_l 是定义在节点上的特征函数, 称为状态特征, 依赖于当前位置. 通常, 特征函数 t_k 和 s_l 取值为 1 或 0. 线性链条件随机域的无向图结构为

$$G = (V = \{1, 2, \dots, n\}, E = \{(i, i + 1)\}) \tag{2}$$

其中, $i = 1, 2, \dots, n - 1$. 在学习阶段, 线性链条件随机域利用训练数据集通过极大似然估计或正则化的极大似然估计得到条件概率模型 $P(Y|X)$; 在预测阶段, 对于给定的输入序列 x , 求出条件概率 $P(Y|X)$ 最大的输出序列 y .

2.2 建立比较要素抽取模型

比较要素抽取任务就是指从比较句中识别出比较实体、比较属性、比较词和评价词等比较成分, 从而获取一个或多个比较关系的五元组 (比较主体, 比较客体, 比较属性, 评价词, 比较词). 本文将比较要素抽取的任务转化为序列标注任务, 通过采用条件随机域模型, 在模型训练过程中引入词语级特征、浅层句法特征、候选比较词及启发式位置特征, 构建了基于条件随机域的比较要素抽取模型. 在整个比较要素抽取任务中, 所选取的模型特征都是领域无关的. 具体的识别过程如图 1 所示.

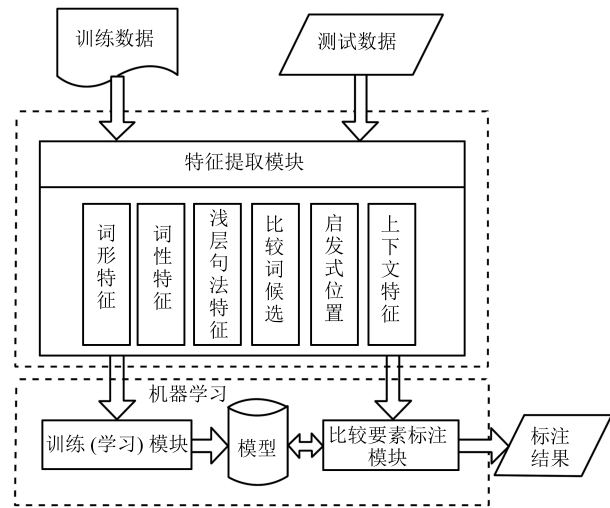


图 1 比较要素抽取的过程

Fig. 1 The procedure of comparative elements extraction

在图 1 中, 识别过程主要分为三个阶段: 特征抽取、模型的学习过程和标注过程. 在特征抽取阶段, 系统抽出训练数据和测试数据的有效特征集; 在学习阶段, 通过设置特征模板选取各种有效特征, 建立起比较要素抽取模型; 在标注阶段, 对测试数据中的比较句进行比较要素的抽取.

3 基于 CRFs 模型的特征集合

3.1 特征构造

3.1.1 浅层句法特征

浅层句法分析也称为部分句法分析或语块划分. 相对于完全句法分析, 浅层句法分析只需要识别句子中结构相对简单的独立成分, 例如非递归的名词短语、动词短语等. 对于给定的比较句, 经过浅层句法分析后, 句子中每个词语都对应一个短语标记. 例如 NP[价格/n] ADV[依然/d] VP[高于/v] DNP[相同/a 配置/n 的/ude] DNP[其他/a 品牌/n 的/ude] NP [汽车/n]. 由于浅层句法分析能够识别短语级的

句法成分, 为自然语言处理任务提供了简单的句法结构信息, 使得它在自动问答、信息抽取等领域中也得到了广泛的应用, 而且取得了比较令人满意的效果. 本文采用的浅层句法分析工具 Stanford Parser 由斯坦福大学开发, 主要的短语定义如表 1 所示.

表 1 浅层句法分析的短语标记
Table 1 The phrase maker of shallow parsing

短语标记	短语类别
NP	名词短语
VP	动词短语
ADJP	形容词短语
PP	介词短语
ADVP	副词短语
ROOT	句子
DNP	所属关系短语
QP	数量词短语
LCP	方位词短语
IP	简单从句

在比较句中, 大部分比较主体和比较属性由名词或名词短语构成, 评价词由形容词或形容词短语构成, 比较词和比较客体通常构成一个介词短语. 在现有的研究中, 将词和词性特征引入特征集合, 只能提取词语级别的比较成分, 而对于短语级别的比较成分则难以正确识别, 因此造成了比较要素识别不准确的问题. 本文提出利用浅层句法分析工具识别名词短语、形容词短语等短语信息, 并将识别的短语以标记的形式作为浅层句法特征引入特征集合.

例如, 在汽车评论中, 比较句“总体来看, 金杯阁瑞斯在动力、环保方面比起风行都略胜一筹”经过浅层句法分析工具的处理, 得到短语句法分析树如图 2 所示.

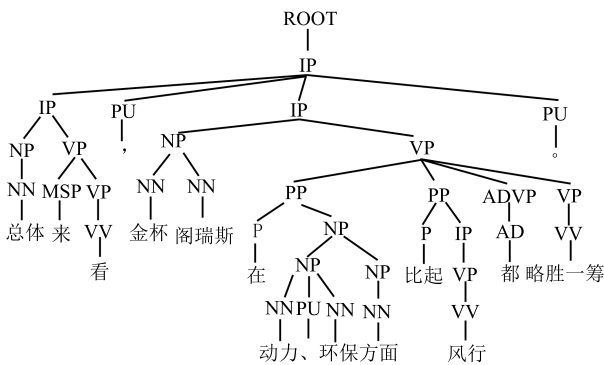


图 2 例句的浅层句法分析树

Fig. 2 The shallow parsing tree of the example

该句的比较主体是“金杯阁瑞斯”, 通过词性标记只能识别出名词“金杯”、“阁瑞斯”, 而无法识别名词短语“金杯阁瑞斯”, 所以引入浅层句法特征有助于识别这类短语要素. 而且, 该句有两个比较属性“动力”、“环保”, 本文通过名词短语首先把两个名词作为一个整体识别为比较属性, 然后利用词语之间的分隔符(或连接词)分离出两个属性. 另外, 介词短语“比起风行”有助于比较词“比起”和比较客体“风行”的识别. 因此, 引入浅层句法特征不仅能够有效识别出短语级别的比较要素, 而且可以作为启发式信息帮助抽取短语中包含的要素.

3.1.2 比较词候选特征

比较词是比较要素识别的重要线索. 为了识别比较词候选, 我们构建了比较词词典, 该词典的构建方法如下: 首先为每种比较类型(包括平比、不同、差比和极比)选择一些常用的种子词语, 然后通过由哈尔滨工业大学信息检索中心开发的同义词词林(扩展版)¹进行同义词扩展, 将扩展后的词语人工剪枝得到比较词词典. 我们尝试使用该词典标注比较词候选, 结果得到了很不准确的结果. 通过对语料的观察发现, 很多比较词是以搭配的形式出现, 例如, “和... 一样”, “与... 不同”, “跟... 相比”. 因此, 我们在比较词词典的基础上构建了比较搭配词典, 该词典采用如下方法构建: 首先统计在每个比较子句中中共现的比较词, 根据其共现频率进行排序, 进而对低频部分进行人工剪枝, 词典中还包含一些常用的比较搭配扩展, 例如: 由比较搭配“和... 一样”扩展出“和(与、跟)... 差不多(相似、相同、无异...)”等搭配, 扩展的基本原理是基于前后搭配项的同义词扩充, 本文共构建了 660 个不同的比较搭配, 通过搭配词典对语料中出现的比较词对做标记, 这些比较词将作为候选比较词.

例如: “单从静态的感觉看, 新飞度总的表现与老款飞度无异”.

“与... 无异”在比较搭配词典中出现, 在其相应的位置做标记, 并作为特征加入特征模板.

3.1.3 启发式位置特征

在比较句的语料中, 大部分主体位于比较词之前或句子(分句)的开端, 客体位于比较词之后或者两个比较词之间. 通过主体、客体与比较词的启发式位置信息, 可以进一步区分比较主体与比较客体. 例如, 句子“奇瑞 A5 的空间比悦动小多了”, 主体“奇瑞 A5”和客体“悦动”分别位于比较词“比”的前面和后面. 但是, 在句子“比起悦动, 奇瑞 A5 的空间小多了”中, 主体“奇瑞 A5”位于比较结构的下一个分句中, 为了有效处理这类句子以及含有多

¹<http://www.ltp-cloud.com/download/#down.cilin>

个比较关系的句子, 我们以分句为单位标注启发式位置信息. 首先, 通过比较搭配词词典标记出分句中出现的候选比较词, 将分句中位于比较词之前的词标识为 LBF, 位于比较词之后的词标识为 LAF, 对于不含比较词的分句, 其所有词标识为 LBF.

例如: “比起悦动, 奇瑞 A5 的空间小多了”.

例句中包含两个子句, 对于第一个子句, 将“比起”标记为候选比较词, 把位于“比起”之后的“悦动”标记为 LAF, 第二个子句不含任何比较词, 因此所有词标记为 LBF.

3.1.4 词形和词性标记特征

词形: 词是自然语言中表达语义的最小单位, 对于西方拼音语言来讲, 词之间有明确的分界符, 而对于中文而言, 词与词之间没有明显的分界符, 因此, 首先需要对句子进行分词, 然后才能使用统计语言模型进行中文信息抽取的工作.

词性标记也是一类重要的特征, 由于比较实体和比较属性主要是名词词性, 评价词主要是形容词或动词词性, 比较关键词一般以介词、副词词性出现, 因此, 引入词性标记作为特征, 将有助于系统区分不同的要素类别.

3.2 条件随机域模型的特征模板

本文采用 CRFs++0.53 开源工具包²进行条件随机域模型的学习和标注工作. 在 CRFs++0.53 中是通过使用特征模板将需要的特征从特征集中提取出来加入到模型中. 因此特征选择的问题就转化成特征模板的选取问题. 本文根据上述语言学相关特征, 设计了如表 2 所示的几种特征组合模板.

表 2 中, w, t 分别代表词、词性特征, c, l 代表比较词候选和启发式位置特征, s 代表浅层句法特征, 为了验证句法特征和启发式特征的作用, 实验中共构建了六种特征模板. 依次是词语级 (Baseline) 特征模板 (T1), 向词语级模板中加入比较词候选特征模板 (T2), 向词特征中加入比较词候选和启发式位置特征模板 (T3), 向词语级模板中加入比较词候选和启发式位置特征模板 (T4), 向词语级模板中加入浅层句法特征模板 (T5) 以及同时加入所有特征时的特征模板 (T6).

4 实验结果及分析

4.1 实验数据

实验数据采用 2012 年第四届中文倾向性分析评测 (COAE 2012) 提供的评测数据 2 作为训练和测试集. COAE 2012 的评测数据 2 包含两个子任务, 子任务 2.1: 从给定的句子集合中, 识别出哪些

句子是比较句; 子任务 2.2: 比较句中相关观点要素的抽取, 从比较句中抽取相应的比较实体、比较属性, 以及对于比较实体的观点倾向性进行判别. 本文的工作与子任务 2.2 相似, 选用的语料来自汽车和电子领域, 共计 1600 个比较句, 其中训练集各包含 200 个比较句, 测试集各包含 600 个比较句, 这些句子大多数是含有比较词的典型比较句, 也包含少量的隐式比较句. 语料的详细统计数据如表 3 所示. 实验中采用的分词工具是中国科学院的 ICTCLAS2013³. 浅层句法分析和依存句法分析工具采用斯坦福大学的 Stanford Parser2012⁴ 实现.

表 2 特征模板
Table 2 Feature templates

模板名称	特征选择	特征模板
T1	w, t	$w_i, t_i, i \in \{-3, \dots, 3\}$ $w_i t_i, i = 0$ $w_{i-1} w_i, t_{i-1} t_i, i \in \{0, 1, 2\}$ $w_i w_{i+1}, t_i t_{i+1}, i \in \{0, 1, 2\}$ $t_{i-1} t_i t_{i+1}, i \in \{-1, 0, 1\}$
T2	w, t, c	$w_i, t_i, c_i, i \in \{-3, \dots, 3\}$ $w_i t_i, t_i c_i, i = 0$ $w_{i-1} w_i, t_{i-1} t_i, c_{i-1} c_i, i \in \{0, 1, 2\}$ $w_i w_{i+1}, t_i t_{i+1}, c_i c_{i+1}, i \in \{0, 1, 2\}$ $t_{i-1} t_i t_{i+1}, c_{i-1} c_i c_{i+1}, i \in \{-1, 0, 1\}$
T3	w, c, l	$w_i, c_i, l_i, i \in \{-3, \dots, 3\}$ $w_i c_i, c_i l_i, i = 0$ $w_{i-1} w_i, c_{i-1} c_i, l_{i-1} l_i, i \in \{0, 1, 2\}$ $w_i w_{i+1}, c_i c_{i+1}, l_i l_{i+1}, i \in \{0, 1, 2\}$ $c_{i-1} c_i c_{i+1}, l_{i-1} l_i l_{i+1}, i \in \{-1, 0, 1\}$
T4	w, t, c, l	$w_i, t_i, c_i, l_i, i \in \{-3, \dots, 3\}$ $w_i t_i, t_i c_i, t_i l_i, c_i l_i, i = 0$ $w_{i-1} w_i, t_{i-1} t_i, c_{i-1} c_i, l_{i-1} l_i, i \in \{0, 1, 2\}$ $w_i w_{i+1}, t_i t_{i+1}, c_i c_{i+1}, l_i l_{i+1}, i \in \{0, 1, 2\}$ $t_{i-1} t_i t_{i+1}, c_{i-1} c_i c_{i+1}, l_{i-1} l_i l_{i+1}, i \in \{-1, 0, 1\}$
T5	w, t, s	$w_i, t_i, s_i, i \in \{-3, \dots, 3\}$ $w_i t_i, t_i s_i, i = 0$ $w_{i-1} w_i, t_{i-1} t_i, s_{i-1} s_i, i \in \{0, 1, 2\}$ $w_i w_{i+1}, t_i t_{i+1}, s_i s_{i+1}, i \in \{0, 1, 2\}$ $t_{i-1} t_i t_{i+1}, s_{i-1} s_i s_{i+1}, i \in \{-1, 0, 1\}$
T6	w, t, c, l, s	$w_i, t_i, c_i, l_i, s_i, i \in \{-3, \dots, 3\}$ $w_i t_i, t_i c_i, t_i l_i, t_i s_i, c_i l_i, c_i s_i, l_i s_i, i = 0$ $w_{i-1} w_i, t_{i-1} t_i, c_{i-1} c_i, l_{i-1} l_i, s_{i-1} s_i,$ $w_i w_{i+1}, t_i t_{i+1}, c_i c_{i+1}, l_i l_{i+1}, s_i s_{i+1},$ $i \in \{0, 1, 2\}$ $t_{i-1} t_i t_{i+1}, c_{i-1} c_i c_{i+1}, l_{i-1} l_i l_{i+1}, s_{i-1} s_i s_{i+1},$ $i \in \{-1, 0, 1\}$

²<http://chasen.org/~taku/index.html.en>

³<http://ictclas.nlp.ir.org/>

⁴<http://nlp.stanford.edu/software/lex-parser.shtml>

表 3 语料规模统计
Table 3 Corpus scale statistics

领域	比较句	比较主体	比较客体	比较属性	比较词	评价词
汽车	800	650	810	836	1 421	831
电子	800	505	860	687	943	802

4.2 评价方法

本文采用信息检索领域标准的三种性能评价指标对抽取结果进行评估: 准确率 (P)、召回率 (R) 和综合评价指标 (F1). 相应的公式表示如下:

$$P = \frac{\text{正确识别出某类比较要素的个数}}{\text{系统输出该类比较要素的总数}} \times 100\%$$

$$R = \frac{\text{正确识别出某类比较要素的个数}}{\text{语料中该类比较要素的总数}} \times 100\%$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

由于比较实体、比较属性和评价词的边界很难明确确定, 不同人的标注也很难统一, 因此本文采用精确评价和覆盖评价两种评价方式. 精确评价是指抽取的实体与答案完全匹配才算正确. 例如: 答案为“屏幕分辨率”, 如果提取结果为“屏幕”或者“分辨率”都不算正确. 覆盖评价是指抽取的实体与答案有重叠就算正确. 例如在上面的例子中, 提取结果为“屏幕”或者“分辨率”都算正确.

4.3 实验结果

4.3.1 比较要素抽取的实验结果

为了获得与在大规模真实文本语料中更接近的实验效果, 我们采用了 5 折交叉验证 (5-fold cross-validation) 的实验方法, 5 次结果的平均值作为对算法精度的估计. 采用上述设计的 6 种特征模板进行实验, 实验结果如表 4 所示, 其中 SUB 代表比较主体, OBJ 代表比较客体, ATTR 代表属性名, KW 代表比较词, SENTI 代表评价词.

表 4 显示了所有比较要素的平均抽取结果, 从实验结果可以发现, 当引入所有特征时 (T6 模板), 系统对各类比较要素的识别结果均优于其他的特征组合策略 (T1 ~ T5 模板). 下面对实验结果进行详细讨论.

从表 4 中可以看到, 使用 T1 特征模板时系统的各项性能指标较差, 尤其是召回率. 分析原因可以发现, 在 T1 模板中仅使用了词形、词性等词语级特征, 而词语级特征所能够提供的信息量是有限的, 而且这些信息中还含有部分噪声. 与之相比, T2 模板将特征范围从词语级扩展到启发式相关信息, 该

模板增加了比较词候选信息, 使各项评测指标均明显提升. T3 模板的性能是两极分化的, 一方面, 在识别比较主体和比较属性时, 获得了最差的性能, 其主要原因在于该模板不包含词性信息, 而词性是比较主体和比较属性的主要指示信息之一; 另一方面, T3 模板增加了比较词候选和启发式位置特征, 使系统识别比较客体、比较词和评价词的性能均有所提升.

表 4 五折交叉验证的平均结果 (%)

Table 4 The average results of 5-fold cross validation (%)

比较要素	特征模板	精确匹配			覆盖匹配		
		准确率	召回率	F1 值	准确率	召回率	F1 值
SUB	T1	67.43	39.03	48.78	74.91	41.53	53.44
	T2	68.47	41.57	50.99	73.29	47.35	57.53
	T3	73.12	32.00	43.83	76.41	36.44	49.35
	T4	70.25	41.81	51.51	75.66	48.01	61.29
	T5	66.08	37.94	48.21	72.25	42.19	53.12
	T6	71.61	41.36	51.54	80.44	50.31	61.90
OBJ	T1	81.60	66.93	73.36	83.00	69.11	75.42
	T2	81.57	69.83	74.99	84.72	72.02	77.86
	T3	78.05	70.63	73.82	78.77	72.71	75.62
	T4	80.75	73.77	76.90	87.88	76.89	82.02
	T5	81.78	66.13	73.02	83.86	68.25	75.25
	T6	82.22	73.03	77.18	91.69	77.21	83.24
ATTR	T1	72.80	48.38	58.13	78.17	50.04	61.02
	T2	74.43	52.83	61.80	79.96	55.57	65.57
	T3	76.51	39.69	52.27	80.66	42.84	55.96
	T4	73.88	52.11	61.11	78.88	55.31	65.03
	T5	71.36	49.71	58.12	75.06	51.73	61.25
	T6	73.70	51.74	60.80	81.95	55.91	66.47
KW	T1	94.63	75.20	83.69	95.64	77.20	85.44
	T2	93.29	85.31	89.05	94.61	86.96	90.62
	T3	92.68	83.99	88.05	93.93	85.59	89.57
	T4	92.87	85.48	88.95	96.92	88.81	92.69
	T5	93.38	77.54	84.68	94.29	79.04	86.82
	T6	93.28	84.70	88.65	96.52	90.42	93.37
SENTI	T1	87.12	61.67	72.17	89.15	62.05	73.17
	T2	87.38	64.55	74.19	88.98	66.48	76.10
	T3	88.69	64.26	74.44	88.77	66.10	75.78
	T4	87.45	68.47	76.74	90.51	68.98	78.29
	T5	86.34	62.95	72.70	88.45	64.97	74.91
	T6	87.08	68.85	76.86	89.30	71.15	79.20

T4 模板在词语级特征基础上增加了比较词候选和启发式位置信息, 提供了候选比较词以及句子中其他成分相对于候选比较词的前后位置信息, 使得各类要素评测指标的召回率和 F1 值均有较大幅度的提升, 这表明比较词候选特征以及启发式位置特征在抽取比较要素的问题上是非常有效的。

但是, T4 模板所提供的特征还具有一定的局限性, 该模板对于短语级要素的识别能力有限。因此 T5 模板将特征范围从词语级扩展到短语级, 该模板增加了浅层句法信息。表 4 显示使用 T5 模板时, 系统对于比较属性、比较词和评价词等要素的 F1 值均有所提高, 但是对于主客体识别的 F1 值有所下降。

最后, 表 4 给出了加入所有特征的 T6 模板对应的最优系统性能的结果。实验结果表明, 与 T1 模板相比, 当引入所有特征 (T6 模板) 后系统对各类比较要素的识别结果的召回率和 F1 值均有大幅度的提升。这也进一步证明了本文提出的各种特征对于比较要素抽取任务是非常有效的。

由于 CRF 是监督学习的方法, 存在领域适应问题, 为了验证本文方法的有效性, 对于汽车评论语料, 我们采用电子领域的语料作为训练集, 得到汽车领域的测试结果, 对于电子评论语料, 采用汽车领域的语料作为训练集, 得到电子领域的测试结果, 两组实验的平均值作为最终实验结果, 如表 5 所示。

表 5 领域交叉标注的结果 (%)

Table 5 The label results of fields crossing (%)

比较要素	特征模板	精确匹配			覆盖匹配		
		准确率	召回率	F1 值	准确率	召回率	F1 值
SUB	T6	58.15	18.51	27.91	64.53	24.74	35.77
OBJ	T6	79.00	58.42	66.65	85.60	61.70	71.71
ATTR	T6	63.53	37.05	45.14	67.75	40.28	50.52
KW	T6	88.42	77.67	82.26	91.55	83.31	87.24
SENTI	T6	80.43	60.27	68.78	82.67	62.59	71.24

将表 5 与表 4 的实验结果进行对比, 我们发现, 表 5 中采用领域交叉训练建立的模型, 其要素抽取性能相比表 4 确实有一定幅度的下降, 其中比较主体和属性的降幅最大, 比较词和评价词的降幅较小, 其原因一方面是由于比较主体、客体和属性是领域相关的, 例如, 比较主体、客体通常是某领域的品牌名称或产品名称, 属性是产品的部件或特性, 因而这三种要素是领域相关的; 另一方面, 比较主体、属性在句子中的位置太过变换, 而且属性不容易和比较主体区别, 因而跨领域标注对主体、属性性能影响最大, 而比较客体由于其位置相对固定, 因而识别性能

稍好于主体和属性。比较词和评价词由于领域相关性不大, 因而其识别性能相对较好。

4.3.2 与 COAE 2012 评测结果的对比分析

我们将系统中采用第 6 类特征模板 (T6) 的要素抽取的平均结果与 COAE 2012 任务 2.2 评测结果最大值的平均值进行了比较。在对比实验中, 仍然采用上述 5 折交叉验证的实验方法, 并取 5 次实验结果的平均值作为最终实验结果的评价值。结果如表 6 所示, 其中 PROD 代表产品名 (比较主体和比较客体), ATTR 代表属性名。

表 6 COAE 2012 数据集实验结果对比 (%)

Table 6 The results contrast on COAE 2012 data (%)

比较成分	方法	精确匹配			覆盖匹配		
		准确率	召回率	F1 值	准确率	召回率	F1 值
ATTR	T6 模板	73.70	51.74	60.80	81.95	55.91	66.47
	COAE_最大值	66.05	62.52	60.78	77.94	67.51	65.69
PROD	T6 模板	76.92	57.20	64.36	86.07	63.76	72.57
	COAE_最大值	67.77	66.05	64.30	82.67	73.58	71.58
PROD+ATTR	T6 模板	75.84	55.38	63.17	84.69	61.14	70.54
	COAE_最大值	60.81	53.89	52.55	67.45	58.56	57.00

表 6 的实验结果显示, 采用本文的方法抽取实体、属性以及实体加属性的结果的 F1 值均高于 COAE 2012 评测结果的最大值, 表明了本文所提出方法的有效性。从实验结果还可以看出, 每类识别结果的准确率都较高, 但是召回率普遍偏低。原因之一是在比较句中主体和属性出现的位置过于灵活, 而且主体常常被省略, 使得模型难以在有限的训练语料中学习普遍的规律, 因此影响了系统的平均召回率。另一个原因是在比较要素识别的过程中, 未引入任何的领域知识, 包括领域知识库、领域词典等。如果在模型训练阶段引入领域词典, 将有助于提高抽取结果的召回率, 但是人工构造领域词典的代价比较大, 而且无法应用于其他领域。因此, 在不引入领域词典的条件下, 想进一步提高系统要素抽取的召回率, 则需要寻找更有效的通用领域的特征来解决这个问题。另外, 系统所有覆盖匹配的评价指标都高于精确匹配的指标, 表明系统对于各类比较要素的定位比较准确, 但是对于要素边界的识别不够准确。其原因主要来自于底层的自然语言处理工具的错误累积, 包括分词、词性标注以及句法分析工具。因此, 底层语言处理技术的提高对于信息抽取准确率的提升具有重要的意义。

4.3.3 实例分析

例 1. 从驾驶的稳定性来看, 别克要比奥迪好很

多.

对于例 1, 采用词语级特征无法获取比较属性“驾驶的稳定性”. 向系统中增加浅层句法特征 (模板 T5) 后能够正确地识别比较属性. 从图 3 的浅层句法分析结果来看, 名词短语“驾驶的稳定性”处于介词“从”和方位词“来看”之间, 这种构成方式在比较句中论述比较属性时比较常见, 因此引入浅层句法特征可以正确识别例 1 中的比较属性.

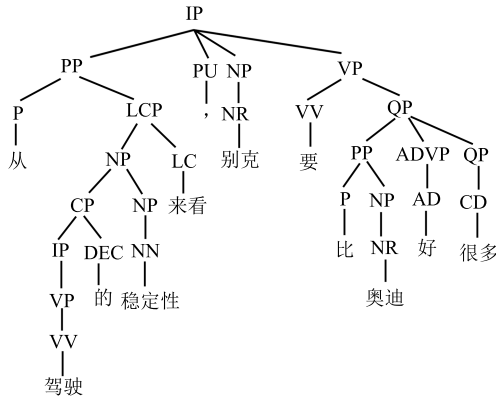


图 3 例 1 的浅层句法分析结果

Fig. 3 The shallow parsing result of Example 1

例 2. 幸福使者配置低内饰和面包车一样.

对于例 2, 使用词语级特征系统能够识别比较主体 (“幸福使者”) 和评价词 (“一样”) 两种要素, 但无法正确识别其他三种要素, 增加浅层句法特征后, 模型正确识别出比较属性 “内饰”、比较词 “和” 以及比较客体 “面包车”. 从图 4 的浅层句法分析结果来看, 句法分析工具引入了部分错误, 例如: “内饰” 应该与动词短语 “和 ... 一样” 构成一个独立分句, 而不是作为前一个分句的一部分 (分句之间没有逗号). 尽管如此, 由 “和” 与 “面包车” 构成的介词短语, 以及名词短语 “内饰” 位于介词短语之前等线索, 模型将词语 “面包车” 识别为比较客体, “内饰” 识别为比较属性.

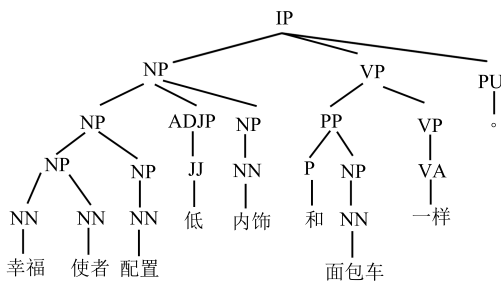


图 4 例 2 的浅层句法分析结果

Fig. 4 The shallow parsing result of Example 2

例 3. 价格依然高于相同配置的其他品牌的汽车 ...

对于例 3, 使用词语级特征只能识别出比较属性 “价格”, 引入比较词候选 “高于” 及启发式位置特征后, 系统识别出了评价词 “高于”, 在引入全部特征后, 系统识别出比较客体 “其他品牌的汽车”. 因此, 该例说明了引入多种特征的有效性.

4.3.4 含有多个比较关系的句子的性能分析

一个比较句中可能含有一个或多个比较关系, 在汽车领域的语料中, 含有多个比较关系的句子占 25.4%, 因此有必要对这类句子的比较要素抽取性能进行分析.

例 4. 从品牌知名度、可靠性、耐用性等角度考虑, 福美来只能算是三流车, 它与宝来、高尔夫、花冠、通用等, 根本不在一个等量级, 这也是为什么福美来只好实行低价格战略的原因.

例 4 中含有多组比较关系, 分别是主体 “福美来” 与客体 “宝来”、“高尔夫”、“花冠” 和 “通用” 之间, 在三个属性 “品牌知名度”、“可靠性”、“耐用性” 上的比较, 例 4 中多个客体以及多个属性之间都是并列关系, 系统能够正确地识别. 该例中只有评价短语 “不在一个等量级” 没有识别出来.

例 5. 百公里油耗在 6.6~7.5 之间, 比老普桑还省一点, 噪音也小.

例 5 中含有两组比较关系, 分别是 “Null” 和 “老普桑” 在 “油耗” 和 “噪音” 上的比较, 此例中系统只识别出了比较词 “比” 和比较客体 “老普桑”, 因为属性 “油耗” 的指示信息很少, 另外, “噪音也小” 与直接评价句形式相同, 因此系统未能识别.

本文对含有多个比较关系的句子的要素抽取的性能进行了评测 (模板 6), 结果见表 7.

表 7 含有多个比较关系的句子要素抽取性能 (%)

Table 7 The performance of element extraction in multi-relation sentences (%)

比较要素	精确匹配			覆盖匹配		
	准确率	召回率	F1 值	准确率	召回率	F1 值
SUB	65.97	43.75	52.61	70.93	44.01	54.32
OBJ	77.62	85.28	81.27	79.42	86.35	82.74
ATTR	76.73	58.54	66.41	78.38	58.90	67.26
KW	92.03	88.62	90.29	92.16	90.48	91.31
SENTI	96.18	64.09	76.92	96.50	65.12	77.76

5 结论

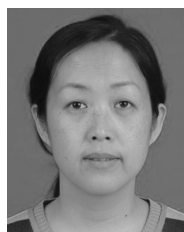
本文面向产品评价领域, 以评价文本中的比较语句为研究对象, 针对比较要素抽取任务, 采用条件随机域模型对引入的多组语言学特征和启发式特征进行融合, 使用有指导学习的方法构建了比较要素抽取模型, 提高了比较关系中各类要素抽取的性能

指标. 实验结果表明, 浅层句法特征能够有效地识别短语级的比较要素, 比较词候选特征不仅能够弥补训练样本中比较词不足的问题, 还可以对其他要素的位置进行初步定位. 启发式位置特征有助于区分主体、客体等词性相似的要素. 模型中引入的所有特征都是领域无关的, 因而该方法可以直接应用于其他领域, 同时对于含有多个比较关系的比较句也能正确处理.

References

- 1 Liu B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (2nd Edition)*. Berlin: Springer-Verlag, 2011. 459–496
- 2 Xu Bing, Zhao Tie-Jun, Wang Shan-Yu, Zheng De-Quan. Extraction of opinion targets based on shallow parsing features. *Acta Automatica Sinica*, 2011, **37**(10): 1241–1247 (徐冰, 赵铁军, 王山雨, 郑德权. 基于浅层句法特征的评价对象抽取研究. *自动化学报*, 2011, **37**(10): 1241–1247)
- 3 Xu K Q. Mining and Analyzing Customer Opinions/Sentiments of Web 2.0 for Business Applications [Ph.D. dissertation], City University of Hong Kong, China, 2011.
- 4 Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: ACL, 2004. 271–278
- 5 Carenini G, Cheung J C K, Pauls A. Multi-document summarization of evaluation text. *Computational Intelligence*, 2013, **29**(4): 545–576
- 6 Jindal N, Liu B. Mining comparative sentences and relations. In: *Proceedings of the 21th National Conference on Artificial Intelligence (AAAI-06)*. Boston, Massachusetts, USA: AAAI, 2006. 1331–1336
- 7 Yang S, Ko Y. Finding relevant features for Korean comparative sentence extraction. *Pattern Recognition Letters*, 2011, **32**(2): 293–296
- 8 Xu K Q, Liao S S, Li J X, Song Y X. Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, 2011, **50**(4): 743–754
- 9 Feldman R, Fresko M, Goldenberg J, Netzer O, Ungar L. Extracting product comparisons from discussion boards. In: *Proceedings of the 7th IEEE International Conference on Data Mining*. Omaha, Nebraska, USA: IEEE, 2007. 469–474
- 10 Xing L Q, Liu L. Chinese standard comparative sentence recognition and extraction research. In: *Proceedings of the 2013 International Conference on Information Engineering and Applications*. Chongqing, China: Springer, 2013. 415–422
- 11 Huang Gao-Hui, Yao Tian-Fang, Liu Quan-Sheng. Mining Chinese comparative sentences and relations based on CRF algorithm. *Application Research of Computers*, 2010, **27**(6): 2061–2064 (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)
- 12 Song Rui, Lin Hong-Fei, Chang Fu-Yang. Chinese comparative sentences identification and comparative relations extraction. *Journal of Chinese Information Processing*, 2009, **23**(2): 102–107 (宋锐, 林鸿飞, 常富洋. 中文比较句识别及比较关系抽取. *中文信息学报*, 2009, **23**(2): 102–107)
- 13 Li S, Lin C Y, Song Y I, Li Z J. Comparable entity mining from comparative questions. *IEEE Transactions on Knowledge and Data Engineering*, 2013, **25**(7): 1498–1509

- 14 Ding X W, Liu B, Zhang L. Entity discovery and assignment for opinion mining applications. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France: ACM, 2009. 1125–1134
- 15 Wang S G, Li H X, Song X L. Automatic semantic role labeling for Chinese comparative sentences based on hybrid patterns. In: *Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence*. Sanya, China: IEEE, 2010. 378–382
- 16 Li Jian-Jun. Research on the Identification of Comparative Sentences and Relations and Its Application [Master dissertation], Chongqing University, China, 2011. (李建军. 比较句与比较关系识别研究及其应用 [硕士学位论文]. 重庆大学, 中国, 2011.)
- 17 Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference Machine Learning*. Williamstown, Australia: Morgan Kaufmann, 2001. 282–289



王 巍 哈尔滨工业大学博士研究生. 主要研究方向为自然语言处理, 情感分析和多文档文摘. 本文通信作者.

E-mail: wangwei@hitwh.edu.cn

(**WANG Wei** Ph.D. candidate at Harbin Institute of Technology. Her research interest covers natural language processing, sentiment analysis, and multi-document summarization. Corresponding author of this paper.)



赵铁军 哈尔滨工业大学教授. 主要研究方向为自然语言处理和人工智能应用.

E-mail: tjzhao@hit.edu.cn

(**ZHAO Tie-Jun** Professor at Harbin Institute of Technology. His research interest covers natural language processing and artificial intelligence application.)



辛国栋 哈尔滨工业大学讲师. 主要研究方向为社交网络和网络安全.

E-mail: gdxin@hit.edu.cn

(**XIN Guo-Dong** Lecturer at Harbin Institute of Technology. His research interest covers social networks and network security.)



徐永东 哈尔滨工业大学博士, 副教授. 主要研究方向为自然语言处理, 多文档文摘和生物信息学.

E-mail: ydxu@insun.hit.edu.cn

(**XU Yong-Dong** Ph.D. and associate professor at Harbin Institute of Technology. His research interest covers natural language processing, multi-document summarization, and bioinformatics.)