面向特定领域产品评价对象自动识别研究

宋晓雷[1] 王素格[1,2] 李红霞[1]

山西大学 数学科学学院, 太原 030006[1],

山西大学计算智能与中文信息处理教育部重点实验室 山西太原 030006[2]

E-mail: wsg@sxu.edu.cn

摘要： 随着 Internet 技术的迅猛发展以及电子商务的不断普及，产品评价对象的识别已成为中文信息处理的一个研究热点。本文首先抽取候选评价对象。通过综合使用词形模板和词性模板以及在对候选评价对象评分之前进行预处理，提高了候选评价对象抽取的召回率和精确率；其次，从模板种子集和评价对象种子集出发，利用自举学习方法对评价对象进行了抽取，并进一步采用 K 均值聚类方法对其聚类，实现了产品名称和产品属性同时自动抽取。实验结果表明，该方法是可行的。

关键词： 评价对象, 模板，K 均值聚类，自举学习

Research on Identifying of Product Evaluation Objects for Specific Domain

Song Xiao-lei[1]， Wang Su-ge[1,2]， Li Hong-xia[1]

Department of Mathematics Science, Shanxi University, Taiyuan 030006,China [1]

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006,China [2]

Abstract: With rapid development of web technology and popularization of e-commerce, the identifying of product evaluating objects is increasingly becoming a hotspot in the Chinese information processing field. Firstly, the candidate evaluation objects are extracted. By combining the word templates with part of speech templates, and preprocessing before giving the scores to the candidate evaluation objects, the recall and precision are improved for extracting candidates evaluating objects. Secondly, from a small set of seeds of templates and evaluating objects, the evaluating objects are extracted by using bootstrapping learning method, furthermore, the evaluating objects are clustered using K-means clustering method for realizing the extracting of product name and product attribute. Experimental results indicate that the approach is feasible.

Keywords: evaluation objects, the templates, K-means clustering, bootstrapping learning

1 引言

随着 Internet 的迅猛发展和电子商务的不断普及，客户评论的数量增长迅速。Web 信息的爆炸使得人名、地名、机构名这三种传统的命名实体识别越来越不能满足人们的需求。产品评价对象的识别日益成为信息科学领域的一个研究热点。在国内，COAE2008 这个首次评测中，20 个国内知名研究机构参与了此次评测，而产品属性的抽取作为 COAE2008[1]的任务之一，其中 13 个单位参加了该项任务的评测，成为参与单位数最多的评测任务；国际上，TREC 会议、ACL 会议也将产品命名实体识别作为其任务之一，众多的国内外单位[2-4]都进行了相关研究。

赵军[2]等在2006年提出了一种基于层级隐马尔可夫模型的产品命名实体识别方法，该方法很

基金项目：国家自然科学基金资助项目(60875040)；教育部高等学校博士点基金(200801080006)；山西省自然科学基金资助项目(2007011042)；教育部科学技术研究重点基金资助项目（2007018）；山西省重点实验室开放基金资助项目；山西高校科技研究开发项目（200611002）

好地处理了多尺度嵌套序列问题；在COAE2008评测的任务3中，文献[5]和文献[6]分别采用最大熵模型和CRF模型取得了不错的成绩；然而采用有监督的学习方法[2,5,6]进行产品命名实体识别时必须获取大量的标注语料，相对来说，需要较多的时间消耗。文献[7-9]都是利用外部资源信息来构造相应的词典，但词典的构建本身是一个难题。本文不需要利用外部信息来构建产品名称、产品属性和情感词词典，只需提供少量产品名称和产品属性，就可自动识别出产品名称和产品属性。Minqing Hu[10]使用关联规则挖掘频繁项作为候选产品命名实体，并对其进行了剪枝处理，然而规则的简单性使其得到产品命名实体识别效果不佳。Hongye Tan[3]等对模板进了泛化，分别使用软模板和特征向量模板对产品命名实体进行了识别，将产品命名实体识别看作分类问题，取得了令人满意的结果；然而文献[3]采用了多领域协同识别，在提高识别性能的同时也限制了它的广泛应用；文献[11-13]采用自举学习方法结合上下文模板进行了英文命名实体识别，然而他们采用的自举方法可以通过在评价候选命名实体之前进行一些简便有效的预处理使其性能得到进一步提高。

总的来说，目前关于产品命名实体识别任务的相关方法还存在以下问题：一是需要大量的基础工作且不便于迁移，代价太大；二是结果不太理想，有待于进一步提高。上述研究都是分别对产品名称或产品属性进行抽取，并没有同时抽取产品名称和产品属性。若能正确地识别出产品名称和产品属性，就可以获取更加详细和精确的产品信息。因此，在没有充分的外部信息的前提下，同时识别出产品名称和产品属性，显得更为重要。

文献[11]的研究表明，特定领域的模板可以极大的提高模板的性能。因此，本文从特定领域展开研究，同时鉴于产品名称和产品属性作为评价对象在语境中具有相似性，在抽取评价对象时采用了同时抽取出产品名称和产品属性的策略。从小种子集出发，综合使用了词形模板和词性模板，通过模糊匹配的方法，提高了候选评价对象的召回率；在评估候选评价对象之前对其进行预处理，提高了候选评价对象的精确率；在进行产品评价对象识别时，采用双向bootstrapping；最后采用K均值聚类进一步对识别结果进行聚类，将其分为产品名称和产品属性。

2 产品评价对象与评价词

（1）产品评价对象

在产品评论中，用户通常关心被评价的对象，但对产品评价对象人们很难给出统一的定义。通过对大量真实的产品评论文本的观察发现，产品评价对象经常是以如下三种方式出现：① 产品的整体；② 产品的某个部件；③产品的特性及其外延。例如：在汽车评论文本中，被评价的对象通常有：*宝马*依旧表现*出色*；速腾的*变速箱*真是*不错*；Polo的*安全*、*质量*和*口碑*也还不错。

为了叙述的方便，本文对上述情形不再细分，我们将第1类的评价对象称为“产品名称”，第2类和第3类的评价对象统称为“产品属性”。

（2）评价词

J.Wiebe[14]的研究表明：形容词可以作为判别句子主客观性的依据，此外，通过大量评论语料观察发现，成语、习惯用语也经常用于评论句。因此，本文选用形容词、成语、习惯用语作为评价词。

3 候选评价对象抽取

通过对大量真实的产品评论文本的观察，我们发现产品评价对象往往是名词或名词短语，苏祺[15]和何婷婷[16]的工作也证明了将名词或名词短语作为候选评价对象是可行的，因此，本文将形式为n、n n、n n n的名词短语作为候选评价对象。

（1）模板的形式（词形模板和词性模板）

模板1：“slot-len,…,slot-i,…,slot-1,word,#”；

模板 2：“#, word, slot+1, …, slot+i, …, slot+len”；

模板 3：“slot-len, …, slot-i, …, slot-1, word, slot+1, …, slot+i, …, slot+len”；

其中：word 表示抽取的评价对象；#表示句子的开始或结束或任意的词或词性；slot-i（slot+i）表示评价对象 word 左面（右面）的第 i 个槽；len 表示窗口的长度。当模板中所有的槽用词形（词性）来表示时，该模板为词形（词性）模板；评价对象与槽可以相邻，也可以不相邻。

例句：“哈飞赛豹 n 的 u 安全性能 n 还是 d 值得 v 信赖 v 的 u”。

由评价对象“哈飞赛豹”从句子中抽取出窗口长度为 1 的词形和词性模板分别为：“#, word, 的”，“#, word, u”。

（2）候选评价对象的抽取

为了获得候选评价对象，本文利用上述模板 1-3，依次搜索评论语料中的每个句子，采用模糊匹配方法对模板与句子进行匹配，仅抽取与模板匹配且距离 slot-1 或 slot+1 最近的名词短语（除去时间、人名、地名、方位名等名词短语）作为候选评价对象，

（3）候选评价对象预处理

为了提高候选评价对象的精确率，在对候选评价对象打分之前对其进行预处理。预处理包括以下三个方面：

<1>去除停用词。这里的停用词包括通用停用词和领域停用词[17]；

<2>中心词剪枝。中心词剪枝采用如下规则：

如果 head(hx)=“车”，则去除 hx 中的中心词。若余下的部分长度大于 1，则将其作为新的候选评价对象，这里的 hx 为候选评价对象。

<3>名词剪枝[9]：有些名词本身并不是商品属性，但它出现在某个商品属性中（例如“高度”与“底盘高度”），而且与该商品属性同时作为候选评价对象被抽取，为了排除此类名词（如“高度”）作为候选评价对象，我们采用规则：如果 $A \subset B$，并且 $count(A) < count(B)$，则去除 A；这里，$count(.)$ 表示某个词或短语在语料中出现的次数。

4 基于 bootstrapping 方法的评价对象抽取

为了获取评价对象，我们采用双向 bootstrapping 方法，其过程为：从小种子集(以模板种子集为例)出发，抽取候选评价对象后，对其采用第 3 节中的方法进行预处理和评分（利用公式（1）进行评分），选取分值最高的前 5 个候选评价对象加入到评价对象集，然后从评价对象集再抽取新的模板(不同于以前的模板)，根据已有的评价对象集对其进行评分，选择分值最高的前 5 个模板加入到模板集，然后再利用现有的模板抽取新的评价对象(不同于以前的评价对象)。重复上述过程，直到没有发现新的符合条件的模板为止。

上述过程中采用的候选评价对象评分标准如下：

$$Score(hx) = \alpha Score_{pjc}(hx) + \beta Score_{c-s}(hx) + \gamma Score_{p-s}(hx) + (1-\alpha-\beta-\gamma) Score_{m-s}(hx) \quad (1)$$

其中：$Score_{pjc}(hx)$ 表示相邻评价词信息，即候选评价对象前后十个位置含有的评价词的数目。$Score_{c-s}(hx)$ 表示词汇(短语)支持度，即词汇或短语在语料中出现的次数。$Score_{p-s}(hx)$ 表示纯支持度[10]，即指候选评价对象作为名词或名词短语在句中出现，并且句中不再包含其它候选评价对象的句子数目。$Score_{m-s}(hx)$ 表示模板支持度，即候选评价对象被模板从语料中抽取出来的次数。本文中，α、β、γ 均取 0.25。

5 产品名称和产品属性的识别

为了把评价对象区分为产品名称和产品属性，本文利用前向选择算法选取文档频率、词频、

段落信息(即候选评价对象在文中的位置信息)三个特征作为聚类特征,进一步采用 k-means 对评价对象进行聚类,其中所用的度量两个向量之间的距离的方法为夹角余弦。

例如,通过对评价对象集中词语聚类,可以找到如下的聚类结果:

{宝马,奥迪,骏捷,思域…};{动力,空间,发动机,内饰…}。

6 实验与分析

(1) 实验数据与评价指标

实验数据采用 COAE2008 的 Dataset2[1]中的汽车评论作为语料库,共有 156 篇评论,平均每篇语料大致包含 6-10 个句子。

评价对象的评价指标:在第 5 节中识别出来的评价对象包括产品名称和产品属性。由于产品评价对象表达形式非常灵活,本文采取了软评测方法[2],并采用三个评价指标:精确率、召回率和 F1。

产品名称和产品属性的评价指标:通过对评价对象聚类,可以得到产品名称与产品属性。本文参考文献[18],采用以下评价指标。其定义如下:

$$precision(T_i,C_j)=n_{ij}/n_j;\qquad precision(T_i)=\max_{C_j\in C} precision(T_i,C_j);$$

$$recall(T_i,C_j)=n_{ij}/n_i;\qquad recall(T_i)=\max_{C_j\in C} recall(T_i,C_j);$$

$$F1(T_i)=2*precision(T_i)*recall(T_i)/(precision(T_i)+recall(T_i))$$

其中:T_i 表示评价对象中应有的某个类别,n_i 表示 T_i 中含有的元素个数,C_j 表示对评价对象聚类所得的某个类别,n_j 表示 C_j 中含有的元素个数,C 表示聚类的总类别。n_{ij} 表示 T_i 与 C_j 共有的元素个数。

(2) 评价对象识别结果与分析

为了识别评价对象,利用第 4 节中基于 bootstrapping 方法进行了以下三个实验。每个实验中的窗口长度均选为 2。

实验 1:验证候选评价对象是否经过预处理对抽取评价对象的影响。

实验 2:验证从不同种子集出发对抽取评价对象的影响。本实验所用的模板均为词形模板。小种子集中种子的个数均选为 7。其中:初始评价对象种子集为:“宝马”,“内饰”,“空间”,“宝来”,“发动机”,“做工”,“奥迪”;初始模板种子集为:“#, word,是,汽车”,“的,word,#”,“#,word,车型”,“#,word,系”,“试驾,word,#”,“#,word,公司”,“#,word,方面”。

实验 3:在实验 2 的基础上,验证词形模板和词性模板及其融合后的模板对抽取评价对象的影响。其中:“A 模型”表示种子集为模板集,其模板为词形模板;“B 模型”表示种子集为评价对象集,其模板为词性模板;“A∪B 模型”表示“A 模型”和“B 模型”的融合模型。

上述三个实验结果见表 1。

由表 1 可知:

<1>每次迭代前对候选评价对象经过预处理比未经过预处理的的效果好,说明对候选评价对象预处理后,一定程度上减少了错误的蔓延,避免了因错误的累积而造成识别性能的急剧下降。

<2>初始种子集为评价对象集或模板集所得的 F1 值相当,且在召回率和精确率上具有一定的互补性。

<3>三个模型中,A 模型和 B 模型在召回率和精确率上是互补的,通过模型的融合,评价对象的 F1 值均有所提高。由于“A 模型”的种子集为我们观察语料所获得的与位置无关的模板集,因此,这些模板的精确率与覆盖率较高,而“B 模型”可能无法召回“A 模型”中某些与位置无

关的模板；但“B 模型”所用的词性模板在某种程度上是词形模板的泛化，因而在某种程度上它们是互补的。

表 1　评价对象识别结果

实验方法 \ 评价指标		精确率	召回率	F 值
实验 1	未预处理	44.84	48.37	46.54
	预处理后	56.88	59.48	58.15
实验 2	评价对象种子集	50.99	67.32	58.02
	模板种子集	56.88	59.48	58.15
实验 3	A 模型	56.88	59.48	58.15
	B 模型	46.35	58.17	51.59
	A∪B 模型	51.75	67.28	58.50

（3）产品名称与产品属性识别结果与分析

为了把评价对象区分为产品名称和产品属性，同时验证我们聚类策略的正确性，采用第 5 节中的 k-means 进行了两个聚类实验：分别对正确的评价对象和上步中识别的评价对象进行聚类。

实验 4：将正确的评价对象聚为产品名称和产品属性两类。

实验 5：直接将通过 bootstrapping 识别的评价对象聚为 2 类，实验窗口长度为 1。

以上两个实验结果见表 2。

表 2　产品名称与产品属性识别结果

实验方法 \ 评价指标	精确率%		召回率%		F1%	
	产品名称	产品属性	产品名称	产品属性	产品名称	产品属性
实验 4	82.86	64.86	87.00	57.14	84.88	60.76
实验 5	73.13	29.73	66.18	25.00	69.48	27.16

由表 2 可知：

<1>实验 4 的结果，对正确的评价对象直接进行聚类时，产品名称和产品属性的 F1 值分别达到了 84.88%和 60.76%。说明本文的聚类方法用于区分产品名称和产品属性是可行的。此外，我们发现实验中识别产品名称的效果显然优于产品属性的效果，主要是由于本文的聚类特征能对产品名称进行很好的描述，因此更倾向于将产品名称聚为一类。

<2>实验 5 与实验 4 相比，实验 5 在识别的性能上下降了很多。主要是由于实验 4 是对正确的评价对象进行聚类，实验 5 是直接对通过 bootstrapping 识别的评价对象进行聚类。通过 bootstrapping 识别的评价对象中不可避免地引入了各种噪声（即非评价对象），使得整体的识别效果不是很理想。

总的来说，上述实验的结果对产品名称的识别是可行的，对于产品属性的识别应进一步研究。

7 结束语

本文给出了特定领域的产品评价对象的定义，提出了一种无监督的学习策略。首先对传统的模板匹配方法进行了改进：综合使用了词形模板和词性模板，在评估候选评价对象之前对其进行预处理；然后，从小种子集出发，识别出产品评价对象后自动对结果进行了聚类，进一步将其分为产品名称和产品属性。整个过程没有用到外部资源，在外部资源不充分的未知领域或新领域处理海量冗余网络数据有一定的指导意义。由于目前还没有同时识别出产品名称和产品属性的相关

实验，我们无法找到已有的研究与我们的实验同时做比较；文献[5]其与位置无关的产品属性抽取的 Lenient 结果的 F 值为 0.1597，我们的 0.2716 与之相比稍高，然而与所有评测结果平均值(与位置无关的 Lenient 结果)：0.49103 相比，我们还有很大的差距。文献[4]采用自举的学习方法结合 HMM 进行英文命名实体识别，在产品名称命名实体(相当于本文的产品名称)识别中获得 69. 18%的 F 值，与本文产品名称识别的 F 值(69. 48%)相近，然而文献[4]的模型复杂度较高；文献[3]在汽车领域的产品名识别中获得 73.1%的 F 值，比本文性能有所提高，但我们的方法有更广的使用范围。此外，我们的方法还有很大的提升空间：（1）可以在后续的研究中改变种子集；（2）聚类中适当添加其它的特征以便减少噪声或者考虑聚为 3 类（产品名称、产品属性以及非评价对象）。

参考文献

[1] 赵军,许洪波,黄萱菁等. 中文倾向性分析评测技术报告[C]// Proceedings of The COAE2008,Harbin,2008:1-20

[2] 刘非凡,赵军,吕碧波等. 面向商务信息抽取的产品评价对象识别研究[J].中文信息学报,2006,20(1):17-20

[3] Hongye Tan,Tiejun Zhao, Jianmin Yao. A Study on Pattern Generalization in Extended Named Entity Recognition[J]. Chinese Journal of Electronic, 2007,16(4):675-678

[4] Cheng Niu,Wei Li, Jihong Ding and Rohini K. Srihari. A Bootstrapping Approach to Named Entity Classification Using Successive Learners[C]// Proceedings of the 41st ACL, Sapporo,Japan,2003:335-342

[5] 何慧,李思,肖芬等. PRIS 中文情感倾向性分析技术报告[C]// Proceedings of the COAE2008, Harbin ,2008:46-55

[6] 张姝,贾文杰,夏迎炬等.基于 CRF 的评价对象抽取技术研究[C]//Proceedings of the COAE2008, Harbin , 2008: 32-37

[7] 王俞霖,孙乐. 软件所 COAE2008 报告[C]// Proceedings of the COAE2008, Harbin ,2008:1-20

[8] 宋锐,林鸿飞. DUTIR 关于 COAE2008 评测报告[C]// Proceedings of the COAE2008, Harbin ,2008:109-114

[9] 赵妍妍,刘鸿宇,秦兵等. HIT_IR_OMS：情感分析系统[C]//Proceedings of the COAE2008, Harbin ,2008:81-88

[10] Mingqing Hu and Bing Liu. Mining and Summarizing Customer Reviews[C]//Proceedings of the tenth ACM SIGKDD.2004:168-177

[11] Etzioni, O., Cafarella, M., Downey, D.,et al. Unsupervised Named-Entity Extraction from the Web: An Experimental Study[J].Artificial Intelligence, 2005,165(1):91-134

[12] Riloff, E., Wiebe, J. and Wilson, T. Learning Subjective Nouns Using Extraction Pattern Bootstrapping[C] // Proceedings of the Seventh Conference on Natural Language Learning, 2003;25-32

[13] Thelen M,Riloff E.A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, USA, 2002:214-221

[14] Wiebe J. , Wilson T. , Bruce R. , Bell M. and Martin M. . Learning Subjective Language [J].Computational Linguistics, 2004, 30(3): 277-308

[15] 苏祺,李芸,王洪俊. 用于产品信息评价的术语库构建及应用[J]. 术语标准化与信息技术.2006(1):33-36

[16] 何婷婷,闻彬,宋乐等.词语情感倾向性识别及观点抽取研究[C]//Proceedings of the COAE2008, Harbin , 2008: 89-93

[17] 黄雄. “小灵通”问答式搜索引擎[R].北京：中科院计算技术研究所,2007

[18] 赵世奇,刘挺,李生. 一种基于主题的文本聚类方法.文信息学报[J].2007, 21(02): 58-62