意见目标网络与意见目标抽取研究*

夏云庆[1] 郝博一[2] 徐睿峰[3]

[1]清华大学清华信息科学与技术国家实验室，北京

[2]清华大学计算机科学与技术系，北京

[3]香港城市大学中文翻译及语言学系，中国香港

E-mail: yqxia@tsinghua.edu.cn; haoby@cslt.riit.tsinghua.edu.cn; ruifeng.xu@cityu.edu.hk

摘　要：未知意见目标是影响意见挖掘系统覆盖率的重要因素。现有意见目标抽取方法大多直接将人工标注的意见目标为种子，通过采取语法/统计模板从真实评价文本中抽取未知意见目标。存在三个问题：（1）手工标注的意见目标粒度过大，不适合作为种子；（2）以列表作为管理种子的数据结构难以表达种子之间的关系；（3）一轮意见目标挖掘往往难以取得满意的效果。该研究提出意见目标网络，采取双层有向图组织原子意见目标和复合意见目标。借助自举策略，意见目标网络能显著提高意见目标抽取的覆盖率。中文意见目标抽取实验结果证明，意见目标网络对处理未知意见目标十分有效。

关键词：自然语言处理，意见挖掘，信息抽取，意见目标抽取

Opinion Target Network for Opinion Target Extraction

Yunqing Xia[1], Boyi Hao[2] and Ruifeng Xu[3]

[1]Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084

[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084

[3]Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong

E-mail: yqxia@tsinghua.edu.cn; haoby@cslt.riit.tsinghua.edu.cn; ruifeng.xu@cityu.edu.hk

Abstract: The unknown opinion targets lead to a low coverage in opinion mining. To deal with this, the previous opinion target extraction methods consider human-compiled opinion targets as seeds and adopt syntactic/statistic patterns to extract new opinion targets. Three problems are notable. 1) Manually compiled opinion targets are too large to be sound seeds. 2) Array that maintains seeds is less effective to represent relations between seeds. 3) Opinion target extraction can hardly achieve a satisfactory performance in merely one cycle. The opinion target network (OTN) is proposed in this paper to organize atom opinion targets (AOT) of component and attribute in a two-layer directed graph. With multiple cycles of OTN construction, a higher coverage of opinion target extraction is achieved via generalization and propagation. Experiments on Chinese opinion target extraction show the OTN is promising in handling the unknown opinion targets.

1　前言

随着网络评价内容的急剧膨胀，意见挖掘研究在自然语言处理研究领域引起了广泛关注。意见挖掘的目的是从评价文本中自动定位并提取相关意见，其难题之一是大量未知意见目标导致意见抽取覆盖率低。研究者试图采取词典和统计相结合的方法自动抽取未知意见目标，典型的做法是：以词典或语料库中人工编辑的意见目标为种子，通过各种扩展方法识别未知意见目标。

Hu 和 Liu(2004)提出借助关联挖掘器获得高频意见目标，以基于意见词的语法模板获得低频意见目标[1]。为了提高覆盖率，Hu 和 Liu（2007）采用了 WordNet 中的同义词集（synset）[2]。

*本文工作得到国家自然科学基金(60703051)、科技部国际科技合作计划(2009DFA12970)、清华大学基础研究基金(JC2007049)和香港城市大学博士后研究基金资助。

Popescu 和 Etzioni(2005)将意见目标视为概念，以“整体-部分”语义关系从评价文本中发现未知意见目标[3]。Ghani 等(2006)以通用的领域无关意见目标作为种子，借助 *co-EM* 算法以“属性–属性值”关系对实现显式和隐式意见目标的推断[4]。Xia 等(2007)借助意见目标与意见关键词的搭配关系寻找未知意见目标[5]。这些工作初步构建了意见目标抽取系统的基本框架，但如下问题仍然十分棘手：(1)人工编撰的意见目标粒度太大，以他们作为种子必然导致覆盖率低的问题。同时，大部分人工编撰的意见目标包含多个词汇，因此无法在同义词集中找到，这导致基于同义词集的方法无法奏效。(2) 种子和意见目标往往都存放在一维数组或列表中，很难有效表示二者关系。例如，多个种子可通过组合形成意见目标。(3) 基于种子的意见目标扩展方法无法一次取得满意的覆盖率，通常需要多次递增式学习才能取得良好效果。

本文提出了意见目标网络（opinion target network, OTN）来解决上述问题。首先，OTN 是一个有向图，其中图的节点（node）代表原子意见目标（atom opinion targets, AOT）同义词集，边（edge）揭示出 AOT 之间的关系，路径（path）则有效地表示了由 AOT 有序组合而成的复合意见目标（compound opinion targets, COT）。在本文的工作中，人工编撰的意见目标被视为 COT，而 AOT 从 COT 中通过泛化自动获得，充当未知意见目标抽取中的种子。由于 AOT 粒度较小，普遍性更强，因此能与同义词集和模式更好配合。其次，OTN 又是一个双层有向图。我们根据本体思想将 AOT 划分为部件（component, COM）和属性（attribute, ATT）。这样，必须采取双层图分别管理部件和属性。本文提出一种自举方法用于构建意见目标网络，经过多次循环，在意见目标网络构建的同时，也能获得评价文本中的未知意见目标。中文意见目标抽取实验结果表明：本文方法在第八轮循环中比基线方法在 f-1 分数上提高了 0.117，在召回率上提高了 0.239。

2 设计原理

2.1 基本思想

本文定义了原子意见目标（AOT）和复合意见目标（COT）。原子意见目标指内聚力强、外在搭配灵活的意见目标。换句话说，从 AOT 内部看，其词汇在统计上彼此依赖很强；而从 AOT 外部看，AOT 能够同很多 AOT 搭配形成真实的意见目标。复合意见目标是指评价文本中以不同模式将 AOT 组合而成的真实意见目标。例如，“图像亮度”是一个复合意见目标，“图像”和“亮度”是原子意见目标。本文工作中，原子意见目标进一步被划分为部件和属性，划分的依据是本体思想。

本文提出的意见目标网络体现了四个设计意图：(1) 该网络能够表示 AOT、COT 以及它们之间的关系；(2) 该网络能区分部件和属性；(3) 该网络能通过同义词集有效表示数万个种子；(4) 该网络可便于以泛化和繁殖方式自动构建。

2.2 形式化表示

意见目标网络（OTN）是一个双层有向图 G^{OTN}，定义为如下五元组：

$$G^{OTN} = < V^{COM}, E^{COM}; V^{ATT}, E^{ATT}; E^{\Theta} >$$

其中 V^{COM} 和 V^{ATT} 分别表示部件节点和属性节点；E^{COM} 和 E^{ATT} 分别表示部件边和属性边；E^{Θ} 则表示部件–属性交叉边。在 OTN 中，路径通常包含了多个边，实际上表示了多个原子意见目标以特定模式有序组合而成的复合意见目标。需要指出：OTN 中的节点都是 AOT 所对应的同义词集，因此一个节点实际上代表了一组 AOT。图 1 给出了 OTN 示例。其中，属性层的虚线边代表了属性分类关系，并不实际用于意见目标抽取。

观察图 1 发现：(1) 在真实评价文本中，属性和部件必须结合在一起才能形成有效意见目标（尽管有时属性可以缺省）。意见目标的核心是属性，它们显式或隐式地与意见关键词搭配组

成意见单元。这一发现让我们了解了意见的形成方法。(2) OTN 中的同义词集和模式能揭示出概念和语义关系，例如整体部分关系。因此我们预测，OTN 可能对自动构建领域本体有帮助。

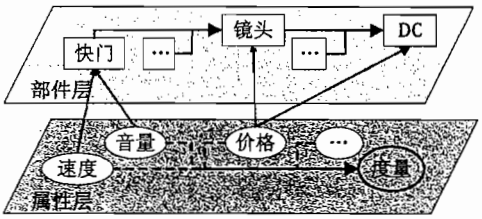

图 1 一个包含部件层和属性层的意见目标网络示例

2.3 构建流程

意见目标网络通过泛化和繁殖多轮自举自动完成，其工作流程如图 2 所示。

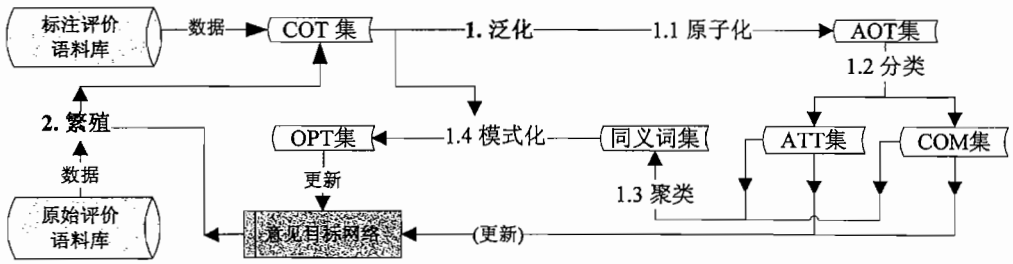

图 2 意见目标网络构建过程。其中，OPT 代表意见目标模式。

图 2 显示：标注评价语料库提供最初的 COT 集。泛化模块从 COT 集中提取 AOT 集，将 AOT 划分为部件和属性，将每个 AOT 赋予特定的同义词集标签，最后形成意见目标模式，形成意见目标网络。繁殖模块借助意见目标网络和依存关系发现未知意见目标。自举算法则执行泛化和繁殖模块，通过多轮学习获得完整的意见目标网络，从而实现高覆盖率的未知意见目标的抽取。

3 泛化过程

3.1 原子化

原子意见目标从大量复合意见目标中以统计方法抽取，主要依据是紧密度和灵活度。紧密度通过计算点式互信息（pointwise mutual information[3]）获得。公式如下：

$$PMI(W_1,W_2)=\ln\frac{P(W_1,W_2)}{P(W_1)P(W_2)}\quad , \tag{1}$$

其中 W_1 和 W_2 是相邻的两个词汇。灵活度计算公式如下：

$$F(W)=\frac{1}{2}\left(\frac{\sum_{W_i\in N^L}\frac{1}{N^R(W_i)}}{N^L(W)}+\frac{\sum_{W_i\in N^R}\frac{1}{N^L(W_i)}}{N^R(W)}\right), \tag{2}$$

其中，N^L 代表左邻词汇集合，N^R 代表右邻词汇集合，函数 N^L(x)返回词汇 x 左邻词汇的类别数目，函数 N^R(x)返回词汇 x 右邻词汇的类别数目。我们设定阈值，选取紧密度和灵活度满足条件的词

汇作为原子意见目标。

3.2 分类

本文设计了一个基于概率的分类器，用于识别部件和属性。该分类器考虑如下两类特征：

1）平均编辑距离（d^{AVG}）

平均编辑距离以字符串编辑距离公式度量将其分类为部件或者属性的概率：

$$d^{AVG}(t \mid X) = \frac{1}{|X|} \sum_{x_i \in X} d(t, x_i), \tag{3}$$

其中，t 表示被分类 AOT，$X=\{x_i\}$ 代表已知部件 AOT 集合或属性 AOT 集合，$|X|$ 表示集合所包含的元素个数，$d(t, x_i)$ 是 t 和 x_i 编辑距离度量函数。根据公式（3），我们可分别度量 t 被分类为部件（C）或属性（A）的概率，并取概率较大者为预测结果。

2）综合位置倾向值（t^{OVA}）

综合位置倾向值利用位置启发信息度量某 AOT 是部件或属性的概率。在某些语言中，位置信息对 AOT 分类具有决定意义。综合位置倾向值计算方法如下：

$$t^{OVA}(t) = \frac{count(t, A)}{count(C, t)}, \tag{4}$$

其中 $count(t, A)$ 是该 AOT t 出现在属性词前的次数，$count(C,t)$ 是该 AOT t 出现在属性词后的次数。

需要指出：初始部件集合和属性集合从标注评价语料库中获得。为提高覆盖率，我们从 HowNet 中抽取更多属性。

3.3 聚类

为了给新发现的 AOT 赋予同义词集标签，我们采取 K-Means 聚类方法将所有 AOT 聚类到适当数目的类簇中。由于 K-Means 聚类方法可通过调节参数获得不同数目的类簇，因此可通过调节参数获得满足如下两个条件的类簇：（1）该类簇包含 3 个以上 AOT，且这些 AOT 同属一个同义词集。（2）该类簇包含至少一个新 AOT。一旦我们找到了这样的类簇，我们将已知 AOT 的同义词集标签赋予所有新 AOT。我们重复上述聚类和赋值过程，直到所有新 AOT 都被赋予了同义词集标签。聚类处理中我们采取了两类特征：（1）原始评价语料库中 AOT 的临近词；（2）新 AOT 和已知 AOT 的编辑距离。

通过上述聚类处理可能无法给所有新 AOT 赋予同义词集标签。这时，我们将所有未获同义词集标签的新 AOT 进行再聚类处理，从而试图获取新的同义词集。如果形成满足如下条件的类簇，则认为发现了新的同义词集：（1）该类簇包含 3 个以上新 AOT。（2）所有 AOT 在原始评价语料库中出现次数都超过 3 次。

在发现了新的同义词集后，我们需要给新发现的同义词集赋予一个标签。我们采取在原始评价语料库中出现次数最多的 AOT 作为标签。上述处理后，仍然会有一些新 AOT 无法获得同义词集标签。我们将这些 AOT 暂时搁置，期望在下一轮聚类处理中参与发现新的同义词集。

3.4 模式化

意见目标模式是以下正则表达式：

$$\{A_c\}*\{string\{B_c\}*\}*,$$

其中，A_c 和 B_c 代表 AOT 的同义词集标签，*string* 代表模式中的字符串常量。举例来说，“<图像>的<颜色>”通过字符串“的”组合了“图像”和“颜色”两个同义词集标签。由于 COT 除了包含 AOT，还包含一些非 AOT 字符，我们用 *string* 常量表示他们。模式是大量 COT 的提炼。

3.5 意见目标网络形成

我们以 AOT 所对应的同义词集为节点，以意见目标模式画边，就建立起一个意见目标网络。注意：如果一个新发现的 AOT 被赋予一个已有的同义词集标签，这时不会在意见目标网络上建立新的节点，而是将这些 AOT 加入该同义词集。

4 繁殖过程

4.1 基于 OTN 的繁殖

意见目标网络具备利用 AOT 和模式进行意见目标推理的能力。换句话说，在意见目标网络中，如果同义词集 A 和 B 之间存在某两个 AOT 所形成的边，那么 A 中的所有 AOT 都有可能和 B 中所有的 AOT 建立同样关系。基于这个假设，我们可借助意见目标网络推理产生候选意见目标。自动推理会导致错误候选，因此必须设计过滤机制排除错误。我们采取序列可信度过滤方法，从原始评价语料库中估计某 AOT 出现在另一个 AOT 之前的概率。给定一个候选意见目标 X，它包含 N 个 AOT，即 $X=\{A_i\}_{i=1,\ldots,N}$，序列可信度计算公式如下：

$$SC(X)=\frac{1}{C_N^2}\sum_{i<j} count(A_i,A_j) \ , \tag{5}$$

其中 $count(A_i, A_j)$表示语料库中 A_i 出现在 A_j 之前的次数。我们设定经验阈值过滤候选意见目标。

4.2 基于依存关系的繁殖

基于模式不能发现新同义词集，制约了意见目标抽取覆盖率。为发现新同义词集，我们考察与已知 AOT 发生依存关系的词汇。为保证基于依存关系的繁殖的可靠性，我们设置如下限制：（1）只考察四类依存关系，即 ATT (修饰)、COO (并列)、QUN(数量)和 DE(的字结构)。（2）候选 AOT 与已知 AOT 邻接，除非他们之间存在连词或“的”。（3）候选 AOT 不是形容词或者代词。实验表明，依存关系对发现新同义词集很有帮助。

5 实验

5.1 实验设置

我们在实验中采用了两个语料库。一个是 Opinmine 语料库[6]，它包含 8,990 个人工标注的关于数码相机的意见。另一个是原始评价语料库，包含了 6,000 篇来自相同领域的用户评价。为对本文方法进行评测，我们将 Opinmine 语料库随机划分为二等份，分别作为训练集和测试集。实验中我们采用精确率（p）、召回率（r）和 f-1 分数（f）对方法进行评测。

本文方法以哈尔滨工业大学提供的 LTP[7]实现中文分词和依存分析，同时我们从中文 HowNet 中手工提取初始属性同义词集。

5.2 实验结果

基线方法：直接以人工编撰的意见目标为种子进行未知意见目标的抽取。为了获取未知意见目标，我们在基线方法中使用了依存分析和意见目标模式。本文方法则以原子意见目标为种子，基于意见目标网络进行多轮自举过程实现未知意见目标抽取。设置基线方法的目的，是要同本文方法进行对比，证明意见目标网络在未知意见目标抽取中的贡献。

本文方法中所涉及阈值设置如下：紧密度阈值设置为 0.001，灵活度阈值设置为 0.333，序列可信度阈值设置为 0.8。这些阈值均为来自数据的经验值。实验结果如图 3 所示。

5.3 讨论

图 3 显示，在第一轮处理后，本文方法在 f-1 分数上超出基线方法 0.051。这时，召回率提高了 0.085，准确率损失了 0.014。可见，本文方法在第一轮处理中就以较小准确率代价取得召回率的较大提高。这表明：原子意见目标的确是更好的种子，基于意见目标网络的抽取方法具有较大潜力。第八轮处理后，本文方法在 f-1 分数上超出基线方法 0.117。这时，召回率提高了 0.239，准确率损失了 0.063。这说明自举过程对意见目标抽取具有的重要贡献。同时我们发现，从第六轮开始，f-1 分数开始收敛，在第八轮基本趋于稳定。这说明自举方法在提升性能上是可收敛。即，总可以通过有限轮自举完成较好性能的意见目标抽取。

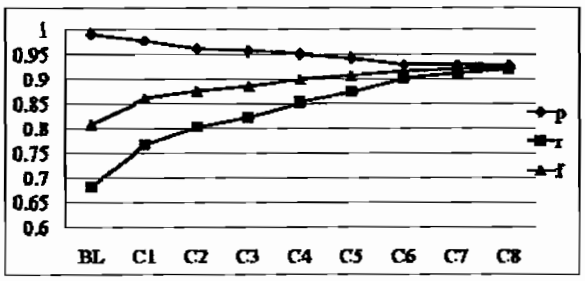

图 3 实验结果。BL 代表基线方法，C1~C8 分别代表本文方法的 8 轮自举过程。

为进一步说明自举过程的贡献，我们观察了每一轮所涉及的 COM、ATT、COT 和模式的个数。统计信息显式：在八轮自举完成后，COM 从 177 个扩展到 1,291 个，ATT 从 67 个扩展到 254 个，COT 从 978 个扩展到 51,724 个，模式从 294 个扩展到 9,077 个。从第六轮开始模式个数逐渐收敛，并在第八轮趋于稳定。这说明自举过程是必要和重要的。

6 结论

本文提出并实现了意见目标网络以提高意见目标抽取的覆盖率。意见目标网络是一个双层有向图，它以原子意见目标（部件和属性）同义词集为节点，通过意见目标模式实现了对复合意见目标的表示。意见目标网络的构建过程恰恰是未知意见目标抽取过程，经过泛化和繁殖的多轮自举处理，显著提高了意见目标抽取覆盖率。本文在中文评价文本上进行了实验，结果表明：意见目标网络对发现未知意见目标具有很大潜力。

参 考 文 献

[1] M. Hu and B. Liu: Mining opinion features in customer reviews. AAAI-2004, pp.755-760 (2004)

[2] M. Hu and B. Liu: Opinion Extraction and Summarization on the Web. AAAI-2006 (2006)

[3] A. Popescu and O. Etzioni: Extracting product features and opinions from reviews. HLT-EMNLP'05, pp. 339-346 (2005)

[4] R. Ghani, K. Probst, Y. Liu, Marko Krema, and Andrew Fano: Text mining for product attribute extraction. SIGKDD Explorations Newsletter, 8(1):41–48 (2006)

[5] Y. Xia, R. Xu, K.-F.Wong, F. Zheng: The Unified Collocation Framework for Opinion Mining. ICMLC-2007. Vol.2, p.844-850 (2007)

[6] R. Xu, Y. Xia and K.-F. Wong: Opinion Annotation in On-line Chinese Product Reviews. LREC-2008 (2008)

[7] J. Ma, Y. Zhang, T. Liu and S. Li: A statistical dependency parser of Chinese under small training data. IJCNLP-04 (2004)

[8] Z. Dong and Q. Dong: HowNet and the Computation of Meaning. World Scientific Publishing (2006)