一种基于 Bootstrapping的本体学习方法

张 俊[1], 高志强[1], 徐 惠[2], 蔡施彦[1], 戴云徽[3]

(1.东南大学 计算机科学与工程学院,江苏 南京 210096; 2.东南大学 软件学院,江苏 南京 210096;
3.南京理工大学 经济管理学院,江苏 南京 210094)

[摘要] 提出了一种基于自扩展的本体学习方法用于获取领域术语.该方法只需提供少量种子术语和一个未标注语料库作为输入,由种子术语开始学习抽取模式,再由学习到的模式发现新的术语,进一步由新发现的术语学习新的抽取模式,如此循环迭代.实验结果表明,该算法能够产生较高质量的领域术语集合和抽取模式集合,这样的集合可用于相关领域的信息抽取.

[关键词] 信息抽取,本体学习,自扩展
[中图分类号] TP 319 [文献标识码] A [文章编号] 1672-1292(2008)04-0056-03

An Ontology Learning Method Based on Bootstrapping

Zhang Jun[1], Gao Zhiqiang[1], Xu Hui[2], Cai Shiyan[1], Dai Yunhui[3]

(1. College of Computer Science and Engineering, Southeast University, Nanjing 210096, China;
2. College of Software Engineering, Southeast University, Nanjing 210096, China;
3. College of Economy, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract This paper presents a bootstrapping-based approach for ontology learning which can be used in field words acquisition. This approach only requires a small set of seed terms and an unmarked corpus as its input, and it learns extraction patterns using seed terms, and detects new field terms using patterns learned before, and again using newly acquired field terms to discover novel extraction patterns, and iterate on. Experiments show that this approach produces a relatively high-quality domain-specific dictionary and a set of extraction patterns which can be consequently applied to information extraction in related domain.

Key words information extraction, ontology learning, bootstrapping

目前,本体已经被广泛应用于语义 Web,信息抽取、数字图书馆[1]等领域.然而构成领域本体的领域术语很难获得,完全手工构建领域术语不仅费时费力,且移植到新领域时需要大量重复劳动.于是提出了自动或半自动构建本体的方法——本体学习.根据 Staab的定义[2],本体学习可分为 6个层次,分别是术语、同义词、概念、分类体系、非分类关系、公理和规则.本文提出的方法主要用于从纯文本中学习领域术语.

1 相关工作

本文提出的基于多模式评分的自扩展算法是一种半监督的本体学习方法.此类方法的特点是根据少量带标注和大量未带标注样例进行学习.其中最具代表性的工作包括 Co-training和元自扩展 (meta-bootstrapping).

Co-training是由 Avrim Blum和 Tom Mitchell提出的用于文本分类的自扩展算法[3].它将文档分为两个视图,让两个视图在迭代学习过程中进行交互,进而改善学习效果.其中,每个页面用(1)页面中文本和(2)其它页面指向该页面的超链接两类信息加以描述,分类器分别使用基于页面的和基于超链接的子分类器进行训练,两子分类器互相迭代改善分类效果.

元自扩展(meta-bootstrapping)方法是由 Ellen Riloff提出的用于从自由文本中获取领域术语的方

收稿日期: 2008-06-18
基金项目: 国家科技计划(2006BAK10B02)资助项目.
通讯联系人: 高志强,教授,博士生导师,研究方向:机器学习和本体学习. E-mail: zqgao@seu.edu.cn

法[4,5]. 该方法使用名短语及其上下文对文档进行术语获取工作, 每轮都对学到的名词短语用启发式方法进行评分, 把高分的样例加到正例集合中去, 该集合可被认为是领域术语集合.

2　基于多模式评分的术语获取方法

本文提出的方法是一种基于自扩展思想的半监督学习方法. 该方法用于从纯文本中自动获取领域术语、生成术语和模式集合.

2.1　模式发现

首先用自然语言处理工具 GATE[6]对训练文档进行解析, 完成对句子的语法和句法成分的标注. 对页面中的每个句子, 用语义词典中的术语对其进行匹配, 根据术语的出现位置决定模式的选取, 最后对获取的模式用 GATE进行人名识别, 同时进行词干化操作, 并用大小为 4的窗口对模式进行截断. 完成以上步骤后, 使用式 1作为模式可信度的评价函数.

$$\mathrm{confidence}(\mathrm{pattern}_i) = \frac{F_i}{N_i} \times \log_2(F_i + 1), \tag{1}$$

其中, $\mathrm{pattern}_i$是第 i个模式; F_i是 $\mathrm{pattern}_i$获取的语义词典词条数; N_i是 $\mathrm{pattern}_i$获取的候选术语总数. 第一项的基础上乘以 $\log_2(F_i + 1)$的目的是增加词条数目在公式中的权重, 以使获得更多词条的模式具有更高的可信度.

2.2　术语获取

在每轮迭代中将最优模式所获取的术语加入到候选术语集合, 以此集合作为选取新的语义词条的基础. 本文使用式 2作为术语的领域相关度评价函数.

$$\mathrm{relevance}(\mathrm{term}_i) = \frac{\sum_{j=1}^{P_i} \log(F_j + 1)}{P_i} + P_i, \tag{2}$$

其中, term_i是第 i个术语; P_i表示获取 term_i的模式数量, F_j表示模式 j获取的语义词典词条数. 对分子取对数降低了单个模式对得分的影响. 同时, 本文认为被更多模式获取的术语更有可能是语义词典的词条, 故将公式加上 P_i.

2.3　算法描述

算法使用种子术语 seedWordSet对语义词典 Lexicon进行初始化. 在每轮迭代中, 将 Lexicon作用于语料库, 使用 2.1节介绍的方法进行模式发现, 使用(1)式计算模式可信度, 将可信度最高的模式放入最优模式集合 Best_pattern_set, 同时将该模式所获取的所有术语放入候选术语集合 Candi_term_set, 对所有在 Candi_term_set中的候选术语使用(2)式计算领域相关度, 选取 10个领域相关度最高且不在 Lexicon中的术语, 完成一轮迭代, 迭代过程重复进行, 直到满足迭代次数 k. 算法描述见表 1.

表 1　基于多模式评分的术语获取算法

Table 1　Term acquisition algorithm based on multi-pattern evaluation

Input　seedWordSet为种子术语集合; Corpus为未标注语料库
Output　Lexicon术语集合; Best_pattern_set存放每轮迭代获取的最佳模式
- tab Initialization: Lexicon ← seedWordSet
- tab while迭代次数 < k do
  - tab new_pattern_set ← 用 Lexicon对页面进行模式发现, 得到新模式集合
  - tab new_pattern_set ← 用 new_pattern_set对页面进行术语获取
  - tab Candi_pattern_set ← Candi_pattern_set+ new_pattern_set
  - tab Best_pattern ← 对 Candi_pattern_set中模式计算可信度, 选出可信度最高者
  - tab Best_pattern_set ← Best_pattern_set+ Best_pattern
  - tab Candi_term_set ← Candi_term_set+ 所有 Best_pattern获取的术语
  - tab Best_term_set ← 计算 Candi_term_set中术语领域相关度, 选出 Top10
  - tab Candi_term_set ← Candi_term_set- Best_term_set
  - tab Lexicon ← Lexicon+ Best_term_set
  - tab return Lexicon and Best_term_set

3　实验结果与分析

本文将该算法应用到 DynamicView系统中,用于学习计算机领域的术语.实验中,共进行 100次迭代.图 1给出了算法性能随迭代次数的变化情况.

由(1)式知,在算法开始迭代阶段,倾向于选取术语获取精度高的模式作为最优模式.而随着迭代次数的增加,语义词典不断扩充,使得对模式的评价更加全局化,这将会引入术语获取精度相对较低的模式.这些模式将会引入部分噪音到语义词典中,噪音反过来又会影响模式可信度的计算,使得算法的精度震荡降低.随着语义词典的不断扩充,召回率上升比较平稳,只是在迭代后期由于噪音的引入变得较为缓慢. F1-Measure值在迭代的大部分时间内也是平稳上升,但在后期随着召回率趋于平稳但精度仍在下降, F1-Measure值略有降低.

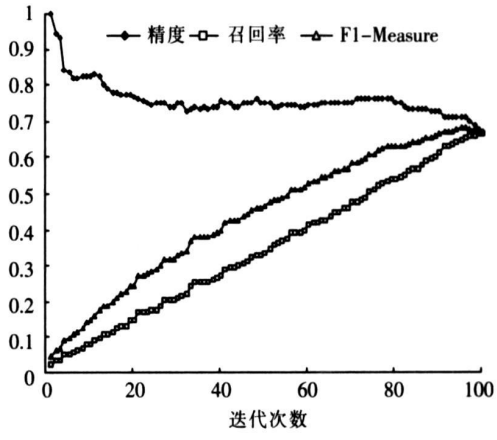

图 1　算法性能随迭代次数的变化

Fig.1　Experimental results in iterative process

4　结语

本文提出了一种基于自扩展的本体学习方法,该方法以一个较小的种子术语集合和一个未标注语料库作为输入,能同时学习领域术语和模式,且具有较高的精度.同时,本文提出的术语获取方法主要是以自然语言处理为基础的,在术语获取领域还存在基于词频、互信息等基于统计技术的方法,将自然语言处理与统计方法相结合将是本文未来工作的重点.

[参考文献](References)

[1] Du X Y, Li M, Wang S. A survey on ontology learning research[J]. Journal of Software, 2006, 17(9): 1837-1847.

[2] Staab S, Hotho A. Semantic web and machine learning[C]// Tutorial at the 22nd International Conference on Machine Learning. Bonn Germany, Amsterdam: IOS Press, 2005.

[3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// COLT. New York: ACM Press, 1998: 92-100.

[4] Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping[C]// Proceedings of the 16th National Conference on Artificial Intelligence. Austin, TX: AAAI Press/The MIT Press, 1999: 1044-1049.

[5] Riloff E. An empirical study of automated dictionary construction for information extraction in three domains[C]// Artificial Intelligence. Karlsruhe: Elsevier Publishers, 1996(85): 101-134.

[6] Cunningham H, Gaizauskas R, Wilks Y. GATE——a general architecture for text engineering[C]// Proceedings of the 16th Conference on Computational Linguistics. Copenhagen, New York: ACM Press, 1996.

[责任编辑:刘　健]