基于 Bootstrapping 的领域词汇自动获取[1]

陈文亮 朱靖波 姚天顺 张宇新

自然语言处理实验室，网络学院

东北大学信息学院计算机软件与理论研究所 辽宁 沈阳 110004

mail:chenwl@mail.neu.edu.cn site:http://www.nlplab.com/

摘 要：领域知识获取是文本处理技术中的一个瓶颈问题，本文提出一种领域词汇的自动获取方法。该方法采用 Bootstrapping 的机器学习技术，从大规模无标注真实语料中，自动获取领域词汇知识。该方法独立于具体领域，移植性好。文中给出了该方法的详细描述。最后，根据实验结果，对该方法的性能进行评估。实验结果显示，从人民日报语料中学习效果比专业领域语料好。

关键词： Bootstrapping, 机器学习，主题识别

Automatic Learning Field Words by Bootstrapping

ChenWenliang ZhuJingbo YaoTianshun

Natural Language Processing Lab

Institute of Computer Software &Theory, Northeastern University, Shenyang 110004

Abstract: This paper presents an automatic learning algorithm that acquires field words. The method is based on bootstrapping. The input to the algorithm is a handful of seed words and unannotated training texts. The method is independent of fields, and it can be applied on many fields for field words learning. In this paper we descript FWB model in detail. According to the experiment, we evaluate the performance of the model. The results present that they are generated from corpus of People's Daily better than corpus of special fields.

Keywords: Bootstrapping, Machine Learning, Topic Identification

1. 前言

主题分析是文本内容分析的一个基础关键技术。朱靖波等[1]曾应用领域知识于文本内容主题分析中，取得了很好的效果。该主题分析方法的性能在很大程度上依赖于一个庞大的领域知识库。1996 年至今，我们已经构造了包含 30 多万项次的领域知识库，主要依靠人工构建，代价十分巨大而且进展缓慢。领域知识库的完善能够大大促进主题分析的效果，但该知识库面临如何扩大规模与增加多语种支持的问题。本文研究的出发点是利用机器学习的方法来试图解决这些问题。

知识获取一直是自然语言处理的重要研究课题。目前很多著名的知识库主要依靠手工构建，如 WordNet[2]，HowNet[3]等。主内题的很多分析是如何文本内取知识容个的题进一大个基础[4][5][6]。关中 Bootstrapping[7][8]键是一技术。朱应用于知识内取的靖波分等技术，Ellen

[1]本文得到国家自然科学基金和微软联合资助项目（60203019）资助

Riloff[9]用来构造信息抽取的知识库，David Yarowsky[10]用来进行语义消歧，等等。

本文提出一种自动获取领域词汇的方法。该方法采用 Bootstrapping 的机器学习技术，从大规模无标注真实语料中，自动获取领域词汇知识。该方法独立于具体领域，移植性好。下文给出该方法的形式化描述。最后，根据实验结果，对该方法的性能进行评估。

2 基本定义

在本文中，有如下重要概念：

a) 领域词 k 是指那些经常出现在特定领域中，且能够表现该领域特征的词，如：金融领域中的证券、股票、金融、人民币、兑付期等词就具备这种特点，是金融领域词。领域词集 F 是领域词的集合，即 k∈K。

b) 在自动学习之前，人工指定一些领域词，本文把这些词称为领域种子词 s。领域种子词集 S 是种子词的集合，即 s∈S。从定义可以看出，s∈K 和 S⊂K。在本文的实验中，各领域的种子词集如表 1 所示，每一个领域选取 10 个种子词。

表 1 实验种子词列表

金融：证券、股票、金融、财经、银行、税收、外汇、投资、股市、贷款
军事：部队、军事、武器、军队、解放军、战争、官兵、军区、核试验、核裁军
体育：体育、选手、运动、足球、中国队、锦标赛、运动员、联赛、决赛、教练
法律：法律、案件、犯罪、执法、法院、法制、违法、律师、司法、检察院

c) 领域重要词 x 是指在 FWB-Model 学习过程中，每一轮获得较高评价的那些领域词。领域重要词集 X 是领域重要词的集合，即 x∈X。

在下文中，“领域种子词”简称为“种子词”，“领域种子词集”为“种子词集”，“领域重要词”为“重要词”，“领域重要词集”为“重要词集”。

3. FWB-Model

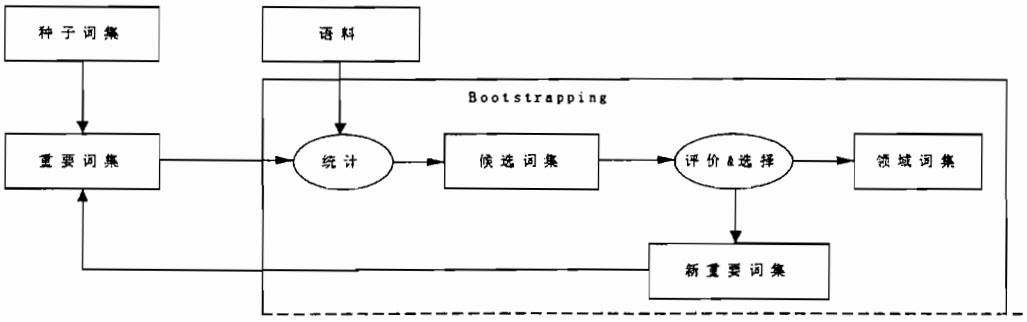

图 1 FWB-Model

本文构建一个基于 Bootstrapping 的领域词汇自动获取模型。该模型从几个种子词出发，在大规模无标注真实语料中，自动获取新的领域词汇。本文把这个学习模型称为

FWB-Model(FieldWords-Bootstrapping)。

整个 FWB-Model 如图 1 所示。FWB-Model 的输入是未标注语料和选定领域的种子词集，学习结果是该领域的领域词集。正如图 1 所示，统计提供了候选词集，而评价与选择是从候选词中选取领域词和领域重要词。FWB 学习算法如算法 1 所示，构造候选词集和评价与选择是算法的核心部分。本文将在 3.1 节描述如何构造候选词集，3.2 节阐述评价与选择部分。

1. 任意 $s \in S$，$s->X$；
2. 从语料中构造候选集 W；
3. 评价与选择
   a) 初评价，从 W 选择符合初评价标准的词加入 W_1；
   b) 再评价，用评价公式进行评价 W_1 中每一个词的 E_w，；
   c) 选择领域词，所有 $E_w >= E_{min}$ 的词为 k_{new}，$k_{new} \in K_{new}$，$K = K \cup K_{new}$ ；
   d) 选择领域重要词，评价值最好的 L 个词为 x_{new}，$x_{new} \in X_{new}$；
4. 如果 $K_{new} = \Phi$，结束学习；
5. $X = X \cup X_{new}$，go to 2；

算法 1 FWB 学习算法

其中，E_{min} 指评价值必须达到的最小值，k_{new} 是新增领域词，K_{new} 是 k_{new} 的集合，x_{new} 是新增领域重要词，X_{new} 是 x_{new} 的集合。

3.1 构造候选词集

本文假设：如果一个词经常和某领域的领域词在一个句子上下文中共现的话，那么这个词也可能是该领域的领域词。所以在本文中，统计单元是一个句子。下列是统计参数的定义：

1. 频数 F_w: 表示在整个语料 C 中包含词 w 的句子数。注意，w 在同一句子中无论出现多少次，都只记为 1 次。
2. 频数 F_X: 表示在整个语料 C 中包含 X 中任意元素 x 的句子数。注意，X 中元素 x 不论在同一句子中出现多少词次，都只记为 1 次。
3. 共现频数 $F_{w,X}$: 表示是 w 与 X 中任意元素 x 在同一句子中共现的句子数。注意，在同一句子不论同时包含多少个 w 和多少个 x，都记为 1 次。

3.2 评价与选择

评价是对候选词集 W 中的每一个词 w 进行评价，得出 w 的评价值 E_w。选择是根据评价结果从 W 中选择合适的词作为新领域词 k，追加到领域词集 K 中，然后从 K 中选择新重要词 x。评价分为两个层次，初评价和再评价。

3.2.1 初评价

初评价的评价条件如下：

1) $F_w >= F_{min}$； 2) $F_{w,X} / F_w >= R_{min}$； 3) $w \notin StopwordList$；

其中，F_{min} 表示必须出现的最少频数，本文把 $R = F_{w,X} / F_w$ 称为支持度，R_{min} 表示支持度的最低阀值，StopwordList 是指禁用词表。所有同时符合这三个条件的词，将构成新候选词集 W_1。

3.2.2 再评价

再评价是对候选词集 W_1 中进行评价，计算词 w 的评价值 E_w。本文使用两种评价公式进行评价，下面分别描述 M 评价、T 评价和(M+T)评价。（下文中所举例子都是金融领域）

3.2.2.1 M 评价

评价公式[9]如下所示：$m_w = Log_2 F(w,X) \times \frac{F(w,X)}{F(w)}$

其中，w 表示词。m_w 值越高，表示 w 是领域词的可能性越大。

把 M 评价应用于 FWB 中，本文通过设定 m_{min} 来评价词 w 是否是领域词，如果 $m_w >= m_{min}$，那么 w 为领域词。

3.2.2.2 T 评价

评价公式[11]如下所示：$t_w = \frac{P(w,X) - P(w)P(X)}{\sqrt{\frac{P(w,X)}{N}}}$

其中 P(w,X)是 w 和 X 同现概率，$P(w,X) = \frac{F_{w,X}}{N}$ (使用最大似然估计来计算概率，以下均同)，P(w)表示 w 出现的概率，P(X)表示 X 出现的概率，N 是句子总数。t_w 值越高，表示 w 是领域词的可能性越大。

使用 T 作为评价指标，应用于 FWB 中，本文通过设定 t_{min} 来评价词 w 是否是领域词。如果 $t_w >= t_{min}$，那么 w 为领域词。

2.2.2.3 M+T 评价

M+T 评价是同时应用 M 评价和 T 评价，在每一轮学习中选取同时符合两种评价标准的词作为领域词，即取两种评价结果的交集。如果 $t_w >= t_{min}$ 和 $m_w >= m_{min}$，那么 w 为领域词。同时，在选择领域重要词时，以 T 评价值为基准。

4. 实验

在实验中采用的是人民日报语料（1994 年、1998 年、2000 年）。首先对语料进行简单分句，也就是利用常见的断句标点（。？！），进行分句。然后，用东北大学自然语言处理实验室的分词工具 CipSegSDK 分词。所有语料总共包含 1,631,540 句，句子平均词数 29.4 词。

实验将在 4 个领域—金融、军事、体育、法律进行，各领域的种子词采用表 1 中所示的种子词列表。学习参数：F_{min}=10，R_{min}=0.5，每一轮选择的新领域重要词最多不能超过 10 个即 L=10。在评价中，本文设定 M 评价中 $m_{min} = Log_2 5 \times (\frac{5}{10} + LoopNum * 0.005)$，其中 LoopNum 是学习的轮次，T 评价中 t_{min}=2.576。

实验 1 语料规模的影响

本实验第一次使用 10 万句，以后每次增加 10 万句，直到加入所有语料。图 2 表示语料规模和正确率之间的关系（横坐标表示语料的规模，纵坐标表示正确率）。图 3 表示新领域词数和规模之间的关系（横坐标表示语料的规模，纵坐标表示新领域词数）。从图中，可以看出

a)随着规模扩大正确率都有不同程度的降低，而学习的词数也不断增加；b)M 和 M+T 的正确率在规模达到 110 万句开始处于一个较稳定状态；c)M 和 M+T 学习词数从 60 万句开始就处于比较稳定的趋势；d)T 的学习词数在 50 万句是突然快速增长，正确率也很快的下降，说明学习处于一个发散状态；e)M+T 的正确率始终比 M 和 T 高。

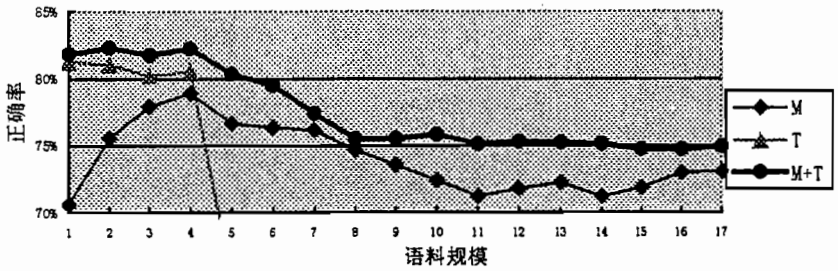

图 2 正确率与语料规模的关系

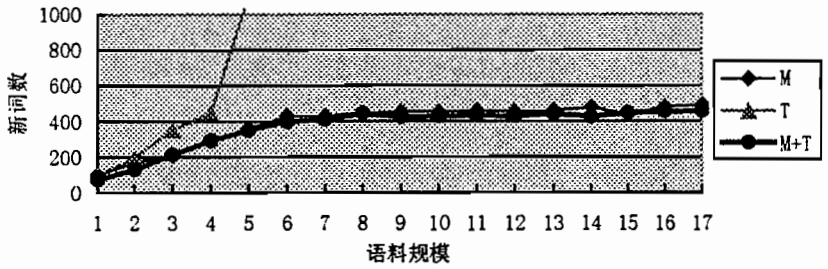

图 3 新词数与语料规模的关系

实验 2 综合评价

表 2 综合评价

领域	评价方式	正确词数/词数	正确率%	领域	评价方式	正确词数/词数	正确率%
金融	M	361/494	73.08	体育	M	150/162	92.59
	T*	665/1000	66.50		T*	867/1000	86.70
	M+T	344/459	74.95		M+T	148/156	94.87
军事	M	179/222	80.63	法律	M	508/705	78.58
	T	197/240	82.08		T*	698/1000	69.80
	M+T	166/199	83.42		M+T	643/781	82.33

本实验对 4 个领域的学习性能进行评估。表 2 中表示学习结果(表中*表示 T 评价学习处于发散状态时，选用 1000 词进行评价)。从实验结果中，可以发现 M+T 评价在所有三种评价中学习效果最好，其中比 M 评价平均提高 2.67%，比 T 评价平均提高 7.62%。

实验 3 专业领域语料的影响

本实验采用各专业语料来评测 FWB 学习性能。金融、军事、体育、法律专业领域语料分别是 20 万句。表 3 表示使用专业领域语料在对应领域所做的实验结果，当实验结果处于发散状态时，本实验选取前 500 词进行评价。实验结果显示，在专业领域语料中 FWB 的学习

效果远不如人民日报。主要原因在于，一个词是否属于领域词汇关键在于它对领域区分的贡献能力，而从专业领域语料中无法获得其他专业领域语料的统计信息，因此基于某一专业领域语料对词的领域区分度进行判断，具有很大的局限性。

表 3 专业领域语料

领域	评价方式	正确词数/词数	正确率%	领域	评价方式	正确词数/词数	正确率%
金融	M	200/500	40.00	体育	M	161/500	32.20
	T	213/500	42.60		T	183/500	36.60
	M+T	214/500	42.80		M+T	183/500	36.60
军事	M	101/500	20.20	法律	M	172/500	34.40
	T	112/500	22.40		T	211/500	42.20
	M+T	108/500	21.60		M+T	217/500	43.40

5. 结论

本文提出一个自动获取领域词汇的学习模型-FWB Model。该学习算法的输入是一些领域种子词和大规模未标注语料，通过 Bootstrapping 学习方法，自动获取领域词。我们发现，词对领域的区分能力是很有限的，所以在领域知识库中，大多数特征项是词串。同时，在实验结果中，我们发现学习到的领域词数目不多，但是学习到的领域词，通过组合可以得到领域特征明显的词串。如：金融领域的“危机”和“金融”组合成“金融危机”、“暴跌”和“股市”组合成“股市暴跌”；体育领域的“悉尼”和“奥运会”组合成“悉尼奥运会”、“海牛”和“队”组合成“海牛队”等等。其中，“危机”、“暴跌”并不是金融领域词，“悉尼”、“海牛”、“队”也不是体育领域词。因此，下一步研究的重点是获取词串来表示领域特征，来扩充领域知识库。我们将使用 FWB 学习到的领域词作为中心词，讨论他们之间甚至多词之间的搭配关系或者他们与其他词之间搭配关系，利用合适的机器学习方法，自动获取领域词串。

参考文献

[1] 朱靖波,姚天顺,基于 FIFA 算法的文本分类,中文信息学报,Vol16,No3,2002

[2] Miller G, WordNet: An On-line Lexical Database. International Journal of Lexicography, 1990

[3] HowNet, http://www.keenage.com

[4] Roman Y, Ralph G, Pasi T, Silja H,Automatic Acquisition of Domain Knowledge for Information Extraction, Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)

[5] 李晓黎,刘继敏,史忠植,概念推理网及其在文本分类中的应用,计算机研究与发展,2000.9

[6] 朱明,林世隆,王俊普,一种聚类型基于示例学习新方法,计算机研究与发展,2000.11

[7] Steven Abney, Bootstrapping, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02) 2002

[8] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In COLT: Proceedings of the Workshop on Computational Learning Theory. 1998

[9] Ellen Riloff, Rosie Jones, Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, Proceedings of the Sixteenth National Conference on Artificial Intelligence(AAAI-99),1999

[10] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), 1995

[11] Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing (141-177), MIT Press, 1999