基于 Bootstrapping 的领域多词串自动获取[1]

郑妍 肖桐 朱靖波

东北大学信息学院自然语言处理实验室 沈阳 110004

E-mail: cindy.zhengyan@163.com

摘 要：在构建领域知识库过程中，领域多词串比词携带了更多的语义信息，对于文本的主题分析和文本的内容分析明显的效果。本文首先利用 C-Value 方法从大规模无标注的真实语料中获取大量的多词串，然后采用 Bootstrapping 的机器学习技术，自动获取多词串的领域特征。实验结果表明，该方法有较好的性能，可以大大减轻人工构建的代价。

关键词：领域知识，机器学习，Bootstrapping，C-Value

Domain Multi-Word Term Acquisition by Bootstrapping

ZhengYan XiaoTong ZhuJingbo

Natural Language Processing Lab, Shenyang 110004

E-mail: cindy.zhengyan@163.com

Abstract: When constructing the domain knowledge base, multi-word term has more powerful represenation than word. In this paper, we present an automaic learning method to extract multi-word term and label its domain feature. First the C-Value method is applied to acquire a number of multi-word terms. Then the proposed method automatically learns domain feature using Bootstrapping. According to the experiment result, Our domain multi-word term acquisition method is effective to reduce human cost for constructing the domain knowledge base.

Keywords: Domain Knowledge, Machine Learning, Bootstrapping, C-Value.

1 引言

研究表明领域知识能有效的改善文本分类的性能[1]、促进主题分析的效果[2]，其应用范围可延伸至众多研究方向[7]，随着信息技术与实际应用的广泛结合，对领域知识库的需求也越来越迫切，知识库的规模更是对应用性能有着直接的影响，但领域知识库的获取方法大多依靠人工构建，其过程繁琐且缓慢，代价巨大。如何利用自动或半自动的方法减轻人工代价是领域知识库构建中的重要问题。

在中文领域知识库的研究中，比较有代表性的是陈文亮[3]提出的自动获取领域词汇的学习模型，并取得了很好的效果，但其候选词集的构建并不充分，且单个的词对领域的区分能力还有其限局性。实际上在真实文本中，存在大量领域特征更为明显的多词串，如体育领域中的“申花”与“主帅”组合，形成多词串“申花主帅”、“鲁能”与“泰山”组合，形成多词串“鲁能泰山”等等，在这些例子中，单个的词不足以表示其领域特征，但多词串作为一个词序列的整体，其领域特征就很明显了。因此本文针对如何构建候选词集及词对领域的区分能力有限的问题，提出以多词串作为基本单元，从大规模无标注的真实语料上自动获取领域多词串及其领域特征。其中，采用 C-Value 方法进行多词串的获取，并在此基础上利用 Bootstrapping 技术自动获取多词串的领域特征。

本文第二节描述领域知识库的定义和组织形式；第三、四节分别描述本文采用的 C-Value 方法与 Bootstrapping 技术的流程与具体算法；第五节描述二组实验的设计与性能的分析，并对实验中遇到的一些问题进行了分析；第六节进行总结和对下一步工作的期望。

[1] 本论文工作得到了国家教育部新世纪优秀人才计划（NCET-05-0287）和国家 863 高科技计划项目（2006AA01Z154）的资助。

2 领域知识库

知识是人类在实践中所积累的认识和经验的总和，领域知识则是与具体领域相关的知识，是人们在日常生活中众所周知的一些动态的知识事实，并在一定程序上体现了领域的发展与变化。如现在提到“姚明”，就会联想到篮球，提到“刘翔”就会联想到田径，不同的领域都包含着大量该领域所特有的知识，很明显，这些领域知识对于人们理解文本有很大的帮助。

领域知识库一般采用分类和分层的方法来进行知识的组织[8]。分类是将领域知识分成不同的领域类别来处理，分层是对同一类别的领域知识分成若干个层次，低层次的概念是高层次的基础，高层次的概念是低层次的概括和总结。图 1 是体育领域的层次分类体系的部分实例。

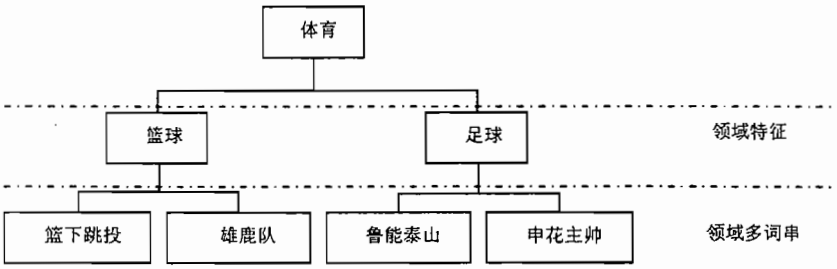

图 1 领域知识库层次分类体系

本文的领域知识库主要关心三方面信息：领域多词串，领域特征，领域属性。如上文提到的领域多词串“鲁能泰山”，它的领域特征为“足球”，它的领域属性为“体育”。本文实验的语料来源于新浪网上大量的体育类新闻，是在已知领域属性的前提下，自动获取领域多词串及其领域特征。

目前，一些较为著名的知识库，如 WordNet[4]，HowNet[5]等，更注重领域知识本体的构建，这些知识库更擅长描述一些通用的信息。其构建过程也更依赖于专家的参与，是一项需要大量精力与时间的工程。

3 多词串的自动获取

本文采用 C-Value 方法[9]从大规模无标注的真实语料中自动获取多词串，该方法将语言学上的知识和统计学上的知识有机的结合起来，并且不依赖于具体领域。

多词串的自动获取流程如图 2 所示：

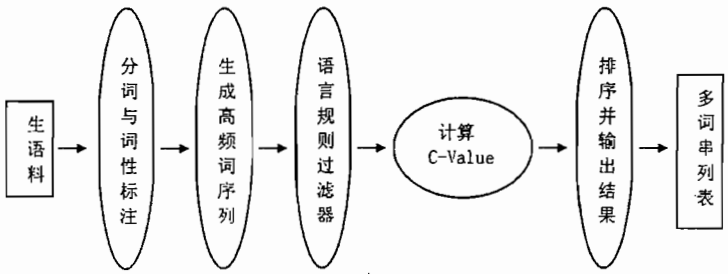

图 2 多词串自动获取流程框图

3.1 候选多词串的自动获取

在文本语料库中，寻找多词串最简单的方法就是基于频率信息的统计方法[10]，如果几个连续

的词经常一起出现，而不是偶然现象，可以假设，它们有固定和特殊的含义，这种含义不能简单解释为两个词的合并。首先要对语料库进行预处理，进行分词与词性标注。利用简单的计数方法统计出语料库中高频的词序列，如果只是考虑统计上的信息，此时会获取大量无用的词序列，如："你\的"，"对\了"等等。本文考虑获取的多词串大多由名词、形容词构成，偶尔带有介词，所以本文通过自己制定一些非常简单的启发式规则，过滤掉不太可能的词序列，来提高结果的准确率。同时，本文也会使用一个禁用词表，因为禁用词表中的词都是在大量文本中出现的高频词，对确定领域特征没有提供太多帮助。本文不希望它出现在获取的多词串中。所以，也需要删除带有禁用词的词序列。

3.2 基于 C-Value 的多词串自动获取

本文选用 C-Value 方法来衡量候选多词串中的词与词之间的相关性有多大，C-Value 方法为每个候选多词串计算一个 C-Value 值，然后按值从大到小的顺序排序，本文从中选取较好的 N 个结果作为多词串。

C-Value 方法是建立于多词串的统计信息基础上的。这些信息包括：

1、候选多词串在整个语料中出现的次数

2、候选多词串作为更长的候选多词串的一部分的次数

3、这类更长的候选多词串的数目

4、候选多词串的长度

对于那些没有被嵌套的候选多词串，它的 C-Value 值只与它在语料中出现的次数和本身的长度有关；对于那些被嵌套的候选多词串，它的 C-Value 值还要考虑它作为嵌套词串的次数，和包含它的更长候选多词串的数目。基于以上信息，本文用如下公式来计算候选多词串的 C-Value 值。

$$C\text{-}Value(a) = \begin{cases} \log_2 |a| f(a) & a \text{ is not nested} \\ \log_2 |a| (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & otherwise \end{cases}$$

这里，a 是候选多词串，b 是更长的候选多词串。$f(a)$是 a 在语料中出现的次数，$f(b)$是 b 在语料中出现的次数，T_a是已经被获取的包含 a 的 b 的集合，$P(T_a)$是集合 T_a中元素的数目。

4 基于 Bootstrapping 的领域特征自动获取

本文采用 Bootstrapping 的机器学习技术[6]，以少量的种子集出发，结合点互信息方法评价元素之间相关与否，多次迭代不断学习，达到自动获取多词串的领域特征的目地。

多词串的领域特征自动获取流程如图 3 所示：

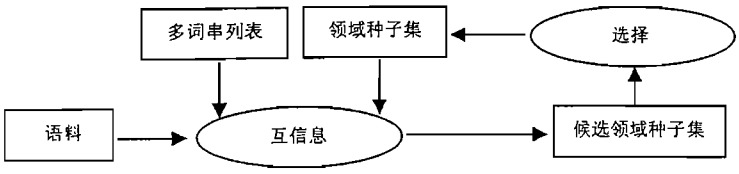

图 3 领域特征自动获取流程框图

4.1Bootstrapping 方法描述

方法的输入是从 C-Value 方法获取的未标注的多词串和领域种子集，学习结果是领域多词串的领域特征。初始状态的领域种子集主要采取人工获取的办法，从 C-Value 方法获取的多词串

列表中，利用常识和经验知识，选择该领域较为有特色和核心的多词串做为初始种子。通过这种方法选择的初始种子，使模型对于任务有较强的适应性。方法描述如下：

1. 人工构造初始领域种子集 S。$x \in S$，$x \in C$；
2. 用互信息方法评价未标多词串 y。$I(x, y)$，$x \in S$，$y \in C$；
3. 生成候选领域种子集。$I(x, y) > -\infty$；
4. 如果候选领域种子集为空，结束学习；
5. 选择领域种子集。选择评价结果最好的前 N 个候选领域种子集生成新领域种子集 S_{new}；
6. $S = S \cup S_{new}$，go to 2；

其中，x、y 代表多词串，S 代表领域种子集，C 代表 C-Value 获取的多词串列表，本文在第 4.2 节中具体描述计算 $I(x, y)$的评价方法。

4.2 基于点互信息的评价

点互信息是一种以信息论为基础的方法，它通常用来计算两个元素之间的互信息，表明元素之间的关联程度。计算公式为：

$$\mathrm{I(x,y)} = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

其中，$p(x, y)$表示元素 x 与元素 y 共同出现的概率，$p(x)$表示元素 x 出现的概率，$p(y)$表示元素 y 出现的概率。本文统计的基本单元为句子，并且采用两种公式进行评价。

4.2.1 多词串与多词串之间（WW）

对于未标多词串 w_i 与领域多词串 w_j 之间计算点互信息 $I(w_i, w_j)$，计算公式如下：

$$\mathrm{I(w_i,w_j)} = \log_2 \frac{p(\mathrm{w_i},\mathrm{w_j})}{p(\mathrm{w_i})p(\mathrm{w_j})}$$

其中，$p(w_i, w_j)$等于句子中共同出现多词串 w_i 与 w_j 的句子数目除以句子总数，$p(w_i)$等于句子中出现多词串 w_i 的句子数目除以句子总数，$p(w_j)$等于句子中出现多词串 w_j 的句子数目除以句子总数。

记录与未标多词串 w_i 的互信息值最大的领域多词串 w_j 的领域特征，此领域特征即为 w_i 的领域特征。此方法需要计算任意两对多词串之间的互信息，且获取的领域特征仅取决于拥有最大互信息时的情况。

4.2.2 多词串与领域种子集之间（WS）

对于未标多词串 w_i 与领域种子集 S_k 之间计算点互信息 $I(w_i, S_k)$，计算公式如下：

$$\mathrm{I(w_i,S_k)} = \frac{\sum_{w_j \in S_k} I(w_i, w_j)}{T(S_k)}$$

其中 $T(S_k)$为领域种子集 S_k 中含有领域多词串 w_j 的数目，且 w_j 是与 w_i 有关联的领域多词串。

对于未标多词串 w_i 与领域种子集 S_k 中的所有领域多词串计算互信息，对互信息求和、求平均。具有最大平均互信息值的领域种子集的领域特征，即为该多词串的领域特征。此方法采用平均互信息的评价方式，一个多词串的领域特征是由领域种子集的平均互信息决定的，而不再由个体独断，能避免偶然的相关性造成的误标，使准确率下降。

进一步，本文对这种方法在效率上进行了优化。本文将语料以类似于倒排索引表的形式表示出来。每个多词串都是一个向量，向量的基是所有的句子，如果多词串在该维度所代表的句子中出现，则用 1 来表示出现，不出现用 0 来表示，不具体计算出现的频次。因此领域种子集 S_k 也可以用一个向量来表示，相当于 S_k 中所有领域多词串向量的内积运算。此时，WS 公式退化为

WW 公式，计算公式如下：

$$I(w_i,S_k)=\log_2\frac{p(w_i,S_k)}{p(w_i)p(S_k)}$$

通过优化，效率大幅度提高，因为本文的领域个数是很有限的。记录与未标多词串 w_i 的互信息值最大的领域种子集 S_k 所属的领域特征，此领域特征即为 w_i 的领域特征。

5 实验

实验采用的语料来源于新浪网的大规模真实的体育类新闻文本。语料库总共包括 2194399 个句子。本文通过 C-Value 方法获取了大量多词串，从中提取前 10000 个具有较高 C-Value 值的多词串进行 Bootstrapping 方法实验。为了对实验性能进行评价，本文人工对这 10000 个多词串标注领域特征，其中人工获取到 4504 个有效的领域多词串，实验性能的评价主要是统计这些有效的领域多词串获取的正确率。在用 Bootstrapping 方法获取领域特征过程中，本文每次从迭代学习到互信息较高的的前 100 个加入到领域种子集中，反复迭代学习，实验共迭代了 100 次完成了获取领域特征的任务。本文的领域特征主要在足球、篮球、排球、乒乓球、游泳、网球、台球、田径、高尔夫、体操这十大类上进行获取。初始领域种子集的选择很重要，本文每个领域都选了 5 个领域多词串做为种子。

5.1 实验设计

5.1.1 多词串获取性能实验

本文采用 C-Value 方法自动获取多词串，获取结果按 C-Value 值从大到小进行了排序。在获取结果中，存在“如“大赛冠军”、“金牌得主”这样的多词串，它可以属于多个领域，该词串的领域特征是由它在语料中的分布决定的，本文将类似于这样的多词串判定为无效数据；而“斯诺克中国公开赛”、“中国选手刘翔”这样的多词串，它所属的领域明显且固定，本文将类似于这样的多词串判定为有效数据。图 4 显示了 C-Value 的有效率与获取结果 TopN 关系。

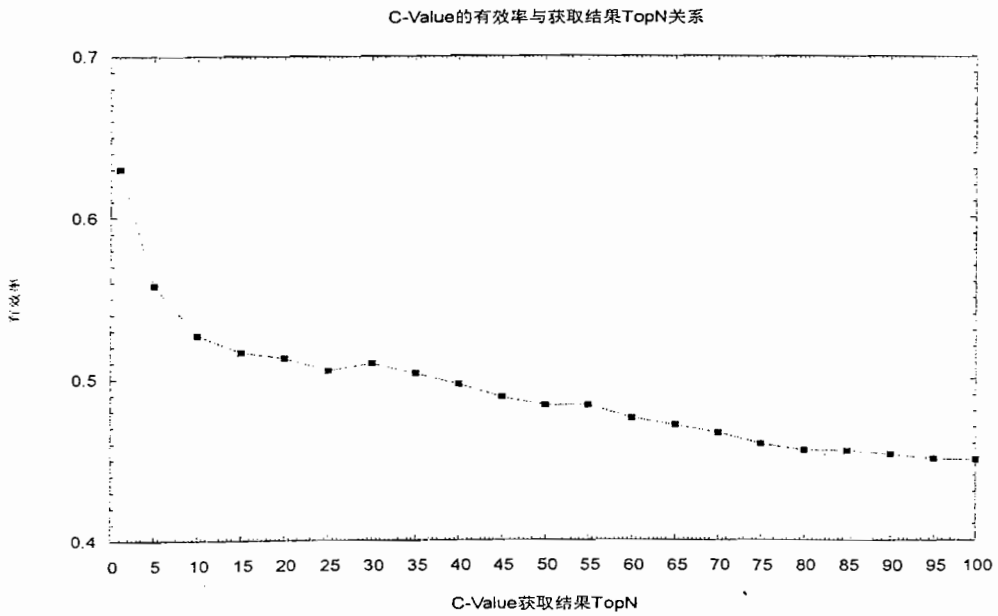

图 4 C-Value 的有效率与获取列表 TopN 关系

图中横坐标表示 C-Value 获取结果 TopN，单位为 100，纵坐标表示有效的多词串所占的比例，总共输出了 10000 个多词串。从图中，本文发现有效数据随着获取规模的扩大而逐渐递减，有效率从 63%下降到 45%。在本文的实验中，最先输出的 2500 个多词串的有效率达到 50%，即大部分多词串是本文所需要的领域多词串。

在本文人工评价过程中，发现数据是否有效还要结合具体任务才能断定，若任务中包含了体育、娱乐、金融等多个领域，则“大赛冠军”、“金牌得主”很明显只属于体育这个领域，则是有效数据。

5.1.2Bootstrapping 标注性能实验

本文采用 Bootstrapping 方法自动标注领域特征，方法从领域较核心的领域多词串出发，每次迭代增加 100 个领域多词串到领域种子集，随着领域种子集的扩大，准确率也必然会受到误标的多词串带来的负面影响。图 5 显示了 Bootstrapping 的正确率与输出结果 TopN 关系。

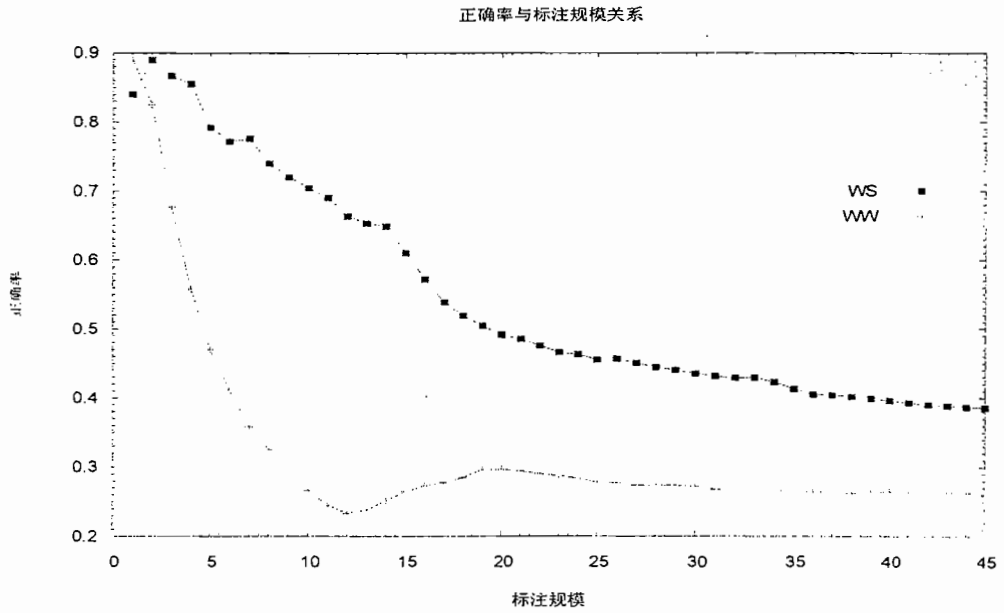

图 5 正确率与标注规模关系

图中横坐标表示 Bootstrapping 获取结果 TopN，单位为 100，纵坐标表示正确获取领域特征的多词串所占的比例，总共输出 10000 个多词串。正如本文所预料的，随着标注规模的扩大，正确率从 91%下降到 38%，达到相对的稳定。Bootstrapping 方法在开始阶段有较好的性能，这是由于还未受于误标的多词串太多影响，但 Bootstrapping 方法会将这种误标引起的错误漫延下去，造到正确率的迅速下降。

图中 WS 与 WW 在第 4.2.1 与 4.2.2 中分别做了分绍。采用 WS 评价的方法性能要明显优于采用 WW 评价的方法性能，WW 的标注结果仅取决于互信息最大时的单个领域多词串，WS 的标注结果则取决于领域种子集的最大平均互信息，判断更具备公平性，避免了语料库中个别样本的不公平信息造成的误标，Bootstrapping 对错误十分的敏感，这也是 WW 的正确率下降的幅度要明显比 WS 多的原因。结合实验的性能与效率两方面考虑，采用 WS 评价的方法的优势更为明显。

5.2 其它问题

初始领域种子集的选择十分重要，首先要避免使用跨领域的多词串，如“接发球”，它即可

以是乒乓球领域的，也可以是网球、排球领域的，这样的种子将使相关领域的标注结果变得混乱。所以应该选择该领域特有的多词串，人名是个很好的选择；其次，要尽量覆盖领域的多个方面，如体操领域又包括自由体操、鞍马，吊环、跳马、双杠、单杠等；最后，本文的标注领域仅限制在十个，但在真实文本中，还存在如滑冰、赛车、奥运等同样重要的领域，如果考虑的比较周全，其正确率也会有很大提高。

实验结果中高频的领域多词串并不一定会比低频的领域多词串先得到，较明显的例子如领域多词串“中国足球”要比“朱广沪”，“主帅福拉多”标注输出位置靠后很多。这也是本文所希望的，最先获取到最具有特色的领域多词串。这也是本文采取互信息计算方法的原因，该方法判断元素之间是否相关要优于判断元素之间的相似程度。

结合本文前面对领域知识库层次体系结构的定义，本文的方法即可以应用在如体育中的足球、篮球、排球、乒乓球等领域特征层次上，也可以应用在如体育、娱乐、金融、军事、教育等领域属性层次上。

6 结论

实验中本文的方法能获取大量领域多词串，如人名，专业术语等，证明了实验的有效性，大大减轻了人工构建代价。本文首先利用 C-Value 方法从大规模无标注的真实语料中获取多词串，然后采用 Bootstrapping 的机器学习技术，自动获取多词串的领域特征。在实验中，Bootstrapping 输出结果位置靠前的多词串，有较高的正确率。但本文的 C-Value 方法只是使用了自定义的简单规则作为语言过滤器，在统计方法上也没有过多的考虑任务本身领域多词串的特性。所以本文下一步研究的重点是如何改进 C-Value 方法或辅助方法，获取更有价值的多词串，期望自动获取领域特征的信任度有更好的提升。

参 考 文 献

[1] 朱靖波，陈文亮，基于领域知识的文本分类，东北大学学报，Vol. 26, No. 8, 2005

[2] 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

[3] 陈文亮，朱靖波，姚天顺，张宇新，基于 Bootstrapping 的领域词汇自动获取，全国第七届计算语言学联合学术会议论文集

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

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

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

[7] 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(COLING2000)

[8] 于楠，朱靖波，陈文亮，领域知识库的构建机制，第二届全国学生计算语言学研讨会论文集

[9] Katerina T.Frantzi, Sophia Ananiadou, Jun-ichi Tsujii, The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms, Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries

[10] Christopher D.Manning, Hinrich Sch tze, Foundations of Statistical Natural Language Processing