文章编号：1003-0077(2010)01-0094-05

基于领域类别信息 C-value 的多词串自动抽取

李 超，王会珍，朱慕华，张 俐，朱靖波

（东北大学 自然语言处理实验室，辽宁 沈阳 110004）

摘 要：该本的多词串抽取是自然语言处理领域一项重要的研究内容。该文提出了一种多类别 C-value(Multi-Class C-value)方法，利用多词串在不同领域的分布信息改善领域相关的多词串抽取的性能。在汽车、科技和旅行三个领域的数据上进行实验，评价多词串的准确率，在 top-100 级别上，较传统的 C-value 方法在三个领域中分别提高了 12、12 和 13 个百分点。实验结果验证了方法的有效性。

关键词：计算机应用；中文信息处理；多词串抽取；多类别 C-value；领域信息

中图分类号：TP391 文献标识码：A

Exploiting Domain Interdependence for Multi-Word Terms Extraction

LI Chao, WANG Huizhen, ZHU Muhua, ZHANG Li, ZHU Jingbo

(Natural Language Processing Lab, Northeastern University, Shenyang, Liaoning 110004, china)

Abstract: Automatic multi-word terms extraction attracts more and more attention in the research of natural language processing. This paper proposes a Multi-Class C-value method, which uses the distribution of multi-word terms in different domains, to improve the performance of multi-word terms extraction. In the experiment with the data of automobile, technology and trip, the precisions of top 100 multi-word terms are 12%, 12% and 13% higher than the clssical C-value method in three domains respectively.

Key words: computer application; Chinese information processing; multi-word terms extraction; Multi-Class C-value; domain information

1 引言

多词串是一种比词携带了更多信息的语言学表示，其应用前景包括信息检索、机器翻译、问答系统、词义消歧以及自动摘要等热门任务[1]。考虑到人工收集多词串的高昂代价以及信息时代领域知识的更新速度，如何用自动或者半自动的方法获取多词串就成为了自然语言处理领域的一个重要问题。

到目前为止，多词串的自动抽取方法包括最开始提出的基于语言学规则的方法[2-4]以及后期提出的基于频率统计的抽取方法[5-8]。C-value 方法是目前用于解决多词串抽取问题最为常用的方法之一。该方法考虑了候选多词串的频次，多词串的长度以及多词串间相互包含的信息并以一个有序的多词串列表作为输出结果。以前的研究工作已经证明了 C-value 方法的有效性[9]。

但是，采用传统的 C-value 方法进行多词串抽取时，对于不同领域的抽取过程是独立进行的，那么就存在一些多词串在多个领域的计算结果中都会得到较大的 C-value 值，最终在输出列表中均获得较高的排位，也就意味着它们在多个领域中同时具有“较高重要性”，对领域类别具有较弱的指示作用，不应该作为领域多词串的抽取结果。

收稿日期：2009-05-25 定稿日期：2009-11-05

基金项目：国家自然科学基金资助项目(60873091)；辽宁省自然科学基金资助项目(20072032)；沈阳市科学技术计划资助项目(1081235-1-00)

作者简介：李超(1986—)，男，硕士，主要研究方向为自然语言处理；王会珍(1980—)，女，博士，助教，主要研究方向为自然语言处理；朱慕华(1981—)，男，博士，主要研究方向为自然语言处理。

本文工作基于如下假设：根据多词串在不同领域的 C-value 输出列表中的位置信息，多词串的排序可以重新调整以获得更好的性能。例如：采用 C-value 方法，多词串“详细 信息”在不同领域的输出列表中都会排在靠前的位置，而多词串“上海 大众”仅在汽车领域中获得较高的排位。在不同领域的 C-value 输出列表中具有相近排位的多词串（例如“详细 信息”）具有较弱的领域相关性，在调整之后的多词串列表中应该赋予较低的排位；与之相对，如果多词串在不同领域输出列表中的位置分布差异较大（例如“上海 大众”），在最终输出结果中应该赋予较高的排位。

基于以上假设，本文提出了一种基于领域类别信息的多词串自动抽取方法：多类别 C-value（Multi-Class C-value）。该方法对 C-value 在不同领域独立输出的结果进行重新排序，得到最终的多词串输出列表。

2　多词串抽取方法

本文提出的多类别 C-value（Multi-Class C-value）方法，首先利用传统的 C-value 计算方法在各个领域中独立进行多词串抽取，生成多词串列表，然后利用各个多词串在不同列表中的位置分布信息进行多词串的重新排序，以获得最终的抽取结果。

2.1　C-value 多词串抽取

使用传统 C-value 方法进行多词串抽取的操作流程可以归纳如下：1）文本预处理，包括分词和词性标注；2）候选多词串的抽取；3）词性规则过滤；4）C-value 值的计算；5）输出各个领域的多词串列表。C-value 的计算方法考虑了候选多词串的长度（词串中词语的个数）、频次信息以及词串相互包含的信息，计算公式如下所示：

$$C\text{-}value(a)=\begin{cases}\log_2|a|\times f(a) & \text{如果 } a \text{ 不被更长的多词串包含}\\ \log_2|a|\left(f(a)-\dfrac{1}{P(T_a)}\sum_{b\in T_a}f(b)\right) & \text{否则}\end{cases}\tag{1}$$

其中：a 表示候选的多词串，$|a|$ 表示多词串的长度，$f(a)$ 表示多词串在整个语料库中出现的频次，T_a 表示以多词串 a 为子串的多词串集合，$P(T_a)$ 表示集合 T_a 中的元素个数。

在利用 C-value 方法进行多词串抽取时，除了 C-value 值的计算，另外一个需要考虑的问题是词性过滤规则的构建。只有符合词性过滤规则的多词串才会参与 C-value 值的计算。由先前工作可知，大部分多词串只由名词、形容词、动词、副词以及介词组成[10]，因此本文所构建的过滤规则只考虑上述五种词性。

2.2　MCC-value 多词串抽取

2.2.1　MCC-value 方法的引入

传统的 C-value 方法在各个领域中分别进行多词串抽取，可以成功地使部分领域相关的多词串在输出列表中排在较高的位置。表 1 给出了在汽车、科技和旅行领域的部分抽取结果。

然而，由于传统的 C-value 方法只考虑了多词串本身在各自领域内的分布信息，而没有考虑多词串在不同领域之间的分布，难以避免会有一部分多词串，在各个领域中都具有较大的 C-value 值，而在最终的输出列表中获得较高的排位，即该类多词串在各个领域中具有类似的分布。直觉上解释，如果某个多词串在各个领域中的分布类似，表示该多词串具有较弱的领域相关性，表 2 显示了部分该类多词串。

表 1　多词串在不同领域的输出列表中的位置情况

多词串	汽车领域	科技领域	旅行领域
最大扭矩	4	42 861	30 566
索尼 爱立信	16 493	28	76 563
风景 名胜区	5 640	16 738	2

表 2　多词串在不同领域的输出列表中的位置情况

多词串	汽车领域	科技领域	旅行领域
知识 产权	22	33	25
有限 公司	13	8	17
详细 信息	3	27	98

在本文中，多词串领域指示性的强弱被称为“领域模糊度”，模糊度的具体计算方法将在 2.2.2 节中详述。

本文提出 MCC-value 方法的动机总结如下：

1）如果多词串 A 只在一个领域的输出列表中排在很靠前的位置，在其他输出列表中没有出现或者是排在很靠后的位置，那么多词串对领域类别具有较强的指示作用，模糊度较低，能够作为领域多词

串的抽取结果;

2) 如果多词串 A 在多个领域的输出列表中都出现在很靠前的位置,那么多词串属于多个领域,对领域类别的指示作用较弱,其模糊度较高,在最终输出的多词串列表中的排位应该被降低。

2.2.2 MCC-value 计算方法

本文利用多词串在传统 C-value 方法输出列表中的位置分布,定义了模糊度计算函数。该函数将被用于对传统 C-value 输出结果进行重新排序,以得到最终的抽取结果。这种考虑多词串在不同领域之间分布信息的 C-value 方法称为多类别 C-value (MCC-value)方法。本文首先定义模糊度计算公式,然后详细介绍如何利用模糊度定义 MCC-value 方法。

某个特定多词串的模糊度(表示为 $AD(t)$)由该多词串在各个领域的输出列表中的位置决定,其计算公式定义如下:

$$AD(t)=\left(\sum_{i=1}^{m}\frac{(\max p(t,S)+1)-p(t,S_i)}{\sum_{j=1}^{m}((\max p(t,S)+1)-p(t,S_j))}\times\log_2\frac{(\max p(t,S)+1)-p(t,S_i)}{\sum_{j=1}^{m}((\max p(t,S)+1)-p(t,S_j))}\right)\Big/\left(\log_2\left(\frac{1}{m}\right)\right) \quad (2)$$

其中:m 表示领域个数,集合 $S=\{S_1, S_2, \cdots, S_m\}$ 表示 C-value 方法得到的 m 个多词串集合,$p(t,S_i)$ 代表多词串 t 在第 i 个领域的 C-value 输出列表中的位置,$\max p(t,S)$ 代表 t 在不同领域的输出列表中位置的最大值,$\log_2(1/m)$ 是归一化因子。公式中的分子部分是一个类似于信息熵的计算式,恰好衡量了多词串在输出列表中的位置差异性,本文称该部分计算式为"位置熵"。

利用公式(2)可以计算得到任意一个多词串的模糊度值。将传统 C-value 方法得到的分值(C-value 值)与 AD 值结合在一起,就可得到基于多类别 C-value 的多词串自动抽取方法,该方法的计算公式定义见公式(3)。

$$MCCvalue(t,S_i)=\log_2 Cvalue(t,S_i)\times(1-AD(t)) \quad (3)$$

其中:$Cvalue(t,S_i)$ 表示多词串 t 在第 i 个领域用传统 C-value 方法计算得到的分值,$AD(t)$ 表示利用公式(2)计算得到的多词串 t 的模糊度。公式中将 $Cvalue(t,S_i)$ 取对数是减弱 $Cvalue(t,S_i)$ 值对于 $MCCvalue(t,S_i)$ 值的影响。由公式(3)可知,模糊度与 MCC-value 的值成反比关系,即模糊度 $AD(t)$ 越小,意味着多词串 t 在多个输出列表中的位置差异性越大,多词串对领域的指示性越强,MCC-value 方法倾向于提高这类多词串的排位。

3 实验

3.1 实验数据

本文采用的语料来自于搜狗语料库 2.0 版本[①]。语料库包含 1 亿个网页。根据对网页的 URL 分析,可以自动得到部分具有领域类别的网页。本文实验采用汽车、科技和旅行三个领域的数据。其包含的网页数量和词的数量见表 3 所示。

表 3 实验数据统计信息

领域	汽车	科技	旅行
#文本	489 830	888 218	324 675
#词	147 058 249	488 805 525	172 472 561

3.2 评价方法

本文采用人工校对的方法对三个领域中的多词串输出列表分别评测。评测的结果采用准确率作为评测指标。为了提高实验结果的可信度,本文的实验结果评测过程中,3 名人员独立进行,并采用了两种评测方法:针对某个抽取方法得到的多词串,评测方法 1,当 3 名评测人员中至少有 2 名人员一致判定多词串抽取结果正确则认为该多词串抽取结果正确;评测方法 2,判定条件更加严格,只有当 3 名评测人员全部判定抽取结果为正确的条件下才认为该多词串的抽取结果为正确。

判定领域多词串正确的基本规则有两个,第一,多词串应该带有明显的领域信息。例如:"上海 大众"多词串携带着明显的汽车领域信息;第二,多词串在语法上必须完整,例如,"责 声明"这样不完整的多词串并不能作为正确的结果。不符合以上两个标准的多词串将判定为错误的结果。

3.3 实验结果

在本文的实验中,候选多词串的最小长度为 2,最大长度设置为 6。由公式(1)可知,除了需要设定

① http://www.sogou.com/labs/dl/t.html

候选多词串的最大长度以外，还需要设定包含当前候选多词串的更长词串的最大长度。在本文实验中该阈值设定为 7。具体地说，假设当前候选多词串 t 的长度为 L，则公式(1)只考虑包含多词串 t 而且长度在[L＋1, 7] 范围内的多词串参与计算。

从表 4 中可以看出 MCC-value 方法较于传统的 C-value 方法有显著的提高。在 top-100 的级别上，使用评测方法 1，汽车领域多词串抽取的结果准确率由 66％提高到 78％，科技领域准确率由 63％提高到 75％，而旅行领域准确率由 64％提高到 77％。在这三个领域中，准确率分别提高了 12％、12％和 13％。随着参与评测的多词串个数增加(最大达到 1 000)，C-value 和 MCC-value 的方法都有明显下降。

表 4　利用评测方法 1 得到的实验结果

汽车领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	66.00	64.00	63.33	61.25	60.60	60.00	60.71	58.38	57.33	55.70
MCC-value/％	78.00	76.00	74.33	73.25	71.80	71.67	68.29	66.75	65.11	64.10
科技领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	63.00	62.00	61.33	61.25	60.20	62.00	62.00	60.88	60.56	59.80
MCC-value/％	75.00	72.50	71.00	69.50	70.40	69.50	68.57	68.13	65.22	64.40
旅行领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	64.00	63.00	62.67	62.00	62.00	62.00	61.00	59.88	59.00	58.30
MCC-value/％	77.00	73.50	71.67	71.50	72.60	70.83	69.14	68.63	67.89	67.40

表 5 给出了使用评测方法 2 得到的实验结果。在 top-100 级别上，准确率由 57％提高到 68％，科技领域准确率由 51％提高到 65％，而旅行领域准确率由 58％提高到 68％。在这三个领域中，准确率分别提高了 11％、14％和 10％。两种评测方法都表明：MCC-value 方法较于传统 C-value 方法有显著提高，充分验证了 MCC-value 方法的有效性。

表 5　利用评测方法 2 得到的实验结果

汽车领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	57.00	53.50	55.00	53.25	52.60	50.50	50.57	48.25	46.11	45.00
MCC-value/％	68.00	67.50	67.33	65.50	61.80	60.50	56.86	54.37	53.00	50.70
科技领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	51.00	45.50	45.00	44.00	42.60	44.50	44.57	43.63	42.89	42.60
MCC-value/％	65.00	60.00	57.33	54.50	55.40	53.50	52.57	51.75	48.89	47.30
旅行领域										
Top-N	100	200	300	400	500	600	700	800	900	1 000
C-value/％	58.00	56.00	52.67	50.75	50.60	49.67	47.43	46.63	45.00	44.30
MCC-value/％	69.00	62.00	59.00	58.50	58.00	55.00	53.29	52.25	50.33	49.60

3.4 实验结果分析

MCC-value 方法,考虑了多词串在不同领域的分布情况,有效地降低了模糊度较高的多词串对于抽取结果的影响,而使用 MCC-value 方法,加入了多词串在不同领域的分布信息,有效地减小了这类模糊多词串对抽取结果的影响。

在多词串抽取结果中,还发现一些错误的多词串抽取结果,例如:汽车领域中的"铅 汽油"、"厢 轿车",这样的多词串在 C-value 方法的输出列表和 MCC-value 方法的输出列表中都排在很靠前的位置,但却不是完整的多词串,不能作为正确的抽取结果。包含"铅 汽油"、"厢 轿车"抽取结果的正确的多词串是"无 铅 汽油"、"两 厢 轿车"这样的多词串,而它们却不符合词性过滤规则,计算 C-value 值时,"铅 汽油"等多词串就会作为不被其他更长的串包含的情况处理,所以影响了抽取结果的准确率。那么词性规则的选择,也一定程度上影响了系统的性能。

4 结论及未来工作

本文首先用 C-value 的方法对多个领域的文本进行多词串自动抽取,然后将多词串在不同领域的分布信息加入到 C-value 方法中,提出了一种多类别 C-value(MCC-value)方法,进行领域多词串的自动抽取。

最后在汽车、科技和旅行三个领域的数据上进行实验,较于传统的 C-value 方法性能有着明显的提高。实验结果表明,此方法在领域多词串自动抽取的任务中是非常有效的。

下一步工作有:1)进一步研究词性过滤规则,寻找更适合于多词串抽取任务的词性规则;2)将 MCC-value 的方法应用到领域知识库的构建工作中,为领域知识库的构建提供多词串信息;3)将模糊度的概念引入到其他多词串抽取方法中,比较其在其他方法中的效果。

参考文献

[1] 段建勇. 多词表达抽取及其应用[D]. 上海交通大学博士论文, 2007.9.

[2] Sophia Ananiadou. Towards a Methodology for Automatic Term Recognition [D]. University of Manchester Institute of Science and Technology, 1988.

[3] Sophia Ananiadou. A methodology for automatic term recognition[C] // Proceedings of the 15th International Conference on Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1994: 1034-1038.

[4] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases [C] // Proceedings of the 14th International Conference on Computational Lingustics. Morristown, NJ, USA: Association for Computational Linguistics, 1992: 977-981.

[5] Ido Dagan, Ken Church. Termight: Identifying and translating technical terminology[C] // Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1994: 34-40.

[6] Beatrice Daille, Eric Gaussier, Jean-Marc Lange. Towards automatic extraction of monolingual and bilingual terminology [C] // Proceedings of the 15th International Conference on Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1994: 515-521.

[7] John S. Justeson, Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identication in text [J]. Natural Language Engineering, 1(1): 9-27, 1995.

[8] Chantal Enguehard, Laurent Pantera. Automatic natural acquisition of a terminology[J]. Journal of Quantitative Linguistics, 1994, 2(1): 27-32.

[9] KT Frantzi, S Ananiadou. The C-Value/NCValue domain independent method for multi-word term extraction [J]. Journal of Natural Language Processing, 1999, 6(3): 145-179.

[10] 朱靖波, 陈文亮. 基于领域知识的文本分类[J]. 东北大学学报, 2005, 26(8): 733-735.