基于粗集理论的中文关键词短语构成规则挖掘

刘远超,王晓龙,徐志明,刘秉权
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 短语比词信息量更加丰富,更能够体现原文的主题,通常所说的关键词实际上多数为短语形式.然而目前的问题是关键词短语的自动标引缺乏统一的规则指导.本文利用粗集理论在数据泛化和知识约简方面的优势,对人工标注的人民日报关键词短语语料进行了挖掘,从而得到了中文关键词短语的若干构成规则.规则可以用于自动关键词抽取,也可以对手工关键词标引进行指导.实验结果表明获取的规则使关键词自动抽取的性能有较大改善.

关键词: 抽取; 关键词短语; 粗集理论; 规则挖掘
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2007) 02-0371-04

Mining Construction Rules of Chinese Keyphrase Based on Rough Set Theory

LIU Yuan-chao, WANG Xiao-long, XU Zhi-ming, LIU Bing-quan
(*Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Phrase conveys more information than word, and can better represent main topic of one article. Most of keywords we referred to are actually in form of phrases. The problem is that extraction of keyphrase lacks guidance of some general rules. By taking advantage of the ability of rough set theory on data generalization and knowledge reduction, the manually labeled keyphrase corpus which come from People's Daily was mined and some construction rules of Chinese keyphrase has been generated. These rules can be used for automatic keyword extraction, and can also help people manually label keyword. The experimental results are promising: the performance of keyword extraction improved greatly after importing these rules.

Key words: keyword extraction; keyphrase; rough set theory; rule mining

1 引言

和文本的标题、摘要一样,关键词提供了可以迅速了解全文信息的重要途径.国外在英文关键词自动抽取方面进行的研究起步较早,如 Eric Brill's Tagger (ftp://ftp.cs.jhu.edu/pub/brill/Programs/) 侧重于抽取频率较高的名词短语作为关键词, NRC's Extractor (http://ai.iit.nrc.ca/II_public/extractor/) 采用有指导的方法学习参数,从而为抽取提供依据.微软公司在其 Office 2000 中也集成了关键词抽取功能.关于中文关键词自动抽取的研究,国内李素建等人[1]对最大熵模型在自动关键词标引任务中的应用进行了有益的分析和探讨,韩客松[2]等人进行了受限范围的主题词标引和主题概念标引,也取得了较好的效果.

通常所说的关键词实际上有相当一部分具有短语形式.短语比词更具有概括能力,包含的信息更加丰富,研究关键词短语的抽取具有重要意义.一种比较常见的研究方法是通过统计 N-gram 词性匹配模式的方法来抽取关键词短语.另外一个相关的研究领域是 Chunk 的自动识别,但 Anette hulth 指出通过 Chunk 自动识别的方法难以获得符合人们习惯的关键词短语,为此她人工总结了 56 个词性匹配模式,用于英文关键词短语的自动抽取[3].

本文工作与 Anette hulth 的相同之处在于挖掘的规则都是基于词性的.区别在于本文挖掘出的规则构成还包括短语左右相邻词的词性.另外由于规则是从足够规模的真实语料中自动获得的,从而使获取的规则更加客观,完备性也可以得到保证.

2 利用粗集理论进行关键词短语的构成规则挖掘

研究表明,只有具有某些词性或词性组合的词和短语才可以被抽取为关键词.关键词短语的构成存在其特殊的规律,但这种规律往往存在于人们的经验之中,而缺乏一个统一的认识.因此有必要获取关键词短语的构成规则,以抽取出符合人们习惯的关键词短语.

为了便于说明问题,首先看几个例子("/"后为该词的词性):

例 1 泰国/ns 政府/nt 13 日/t 决定/v 继续/v 履行/v 其/r 在/p 东盟/ns 贸易区/n 内/f 所/u 承担/v 的/u 义务/n ./w

例 2 中国/ns 民乐/n 除夕/t 回荡/v 维也纳/ns ./w

收稿日期:2006-01-04;修回日期:2006-07-20
基金项目:国家自然科学基金(No.60435020);教育部微软语言语音重点实验室基金(No.01307620)

例 3 阿尔巴尼亚/ ns 高息/ n 集资/ vn 引发/ v 金融/ n 风潮/ n ./ w

例 4 1998 年/ t 将/ d 是/ v 积极/ ad 发展/ v 两岸/ n 关系/ n 的/ u 好/ a 时机/ n ./ w

通过分析,可以初步确定例子中哪些词序列符合关键词短语的构成规律.如例 1 中的"泰国/ ns 政府/ nt "和"东盟/ ns 贸易区/ n "、例 2 中的"中国/ ns 民乐/ n "、例 3 中的"高息/ n 集资/ vn "和"金融/ n 风潮/ n "、例 4 中的"两岸/ n 关系/ n "等.考察一个词序列是否可以作为关键词短语,除了和构成该词序列的词的词性有关外,还和与该词序列相邻的词的词性有关.如例 3 中的"阿尔巴尼亚/ ns 高息/ n "是不应该被抽取为关键词短语的,否则将使"高息集资 "的信息发生割裂.与此类似,例 2 中的"民乐/ n 除夕/ t "也不适合被抽取.

通过对类似上面例子的句子进行分析,我们可以将问题域映射到一个类似表 1 这样的决策表中.表 1 中 w_1 和 w_2 为当前要考察的词的序列,w_L 和 w_R 为这个词序列的左右相邻词.如果没有相邻词则将相邻词设置为句号". / w ".p(w) 表示词的词性,$k(w_1w_2)$ 表示词序列 w_1w_2 是否可以被抽取为关键词短语.$k(w_1w_2)=1$ 表示可以,否则为不可以.由于关键词短语最多可以为 3 个词,所以还需要构造由 3 个词构成的序列的情况,其决策表形式与表 1 类似,只不过需要增加属性 w_3 和 p(w_3).表 1 中 p(*)为条件属性,其取值采用北京大学词性标注集[4].在进行标注时可以提供给标注者相对比较完整的信息(如 w_L、w_1、w_2、w_R 等的取值,以及其词性等),用于标注决策属性的取值.

语料中往往存在较多冗余的知识,需要进行约简.而粗集理论[5]在处理这些问题上则具有较好的优势.由于人工判断本质上主要是根据词性信息,因此只需将相关词的词性信息和 k(*)的取值分别作为粗集的条件属性和决策属性.这里一个重要的工作是如何自动生成具有类似表 1 形式的决策表的条件属性部分,本文采用算法 1 来完成这一工作.

表 1 关键词短语构成规则的决策表(2 个词构成的序列)

序号	w_L	w_1	w_2	w_R	$p(w_L)$	$p(w_1)$	$p(w_2)$	$p(w_R)$	$k(w_1w_2)$
1	在	东盟	贸易区	内	p	ns	n	f	1
2	.	泰国	政府	13 日	w	ns	nt	t	1
3	.	阿尔巴尼亚	高息	集资	w	ns	n	vn	0
4	阿尔巴尼亚	高息	集资	引发	ns	n	vn	v	1
5	.	中国	民乐	除夕	w	ns	n	t	1
6	发展	两岸	关系	的	v	n	n	u	1

算法 1 给定 $n(n=2,3)$ 以及由汉语文本组成的语料库 C_p,构造决策表 I_n

(1) Begin;

(2) 初始化 m、n、t,初始化对象 o,初始化位置索引 P_w, P_p;

(3) 读入语料库 C_p,分词、词性标注,获得词向量 V_w 和词性向量 V_p,且$|V_w|=|V_p|$;

(4) 将 P_w 指向 V_w 的起始位置,将 P_p 指向 V_p 的起始位置;

(5) While $P_w<|V_w|-(m+n+t)$;

(6) Do;

(7) $P_w'=P_w$, $P_p'=P_p$;

(8) For $i=0$ to $2(m+n+t)-1$ //获取决策表中的每个对象(词以及词性);

(i) $o(i)=V_w(P_w')$, $o(i+1)=V_p(P_p')$;

(ii) $i=i+2$;

(iii) $P_w'=P_w'=P_p'+1$;

(9) End for;

(10) 将对象 o 插入到决策表 I_n,作为其中的一行;

(11) 置空对象 o;

(12) $P_w=P_w+1$, $P_p=P_p+1$;

(13) Done;

(14) End while;

(15) End.

算法 1 中| *|的含义是 *中元素的个数,m、n、t 的取值含义为长度为 n 的词序列,其左相邻词和右相邻词的个数分别为 m 和 t. n 的取值可以为 2 或 3,m 和 t 的取值为 1.通过算法 1 获取的决策表规模较大,而且有些对象明显不可能是关键词短语,因此还需要进行适当的过滤处理.定义过滤规则为,如果词序列中任何一个词的属性不具有类名词的特征,则过滤掉此序列.这里所说的类名词的词性包括"n(名词) "、"nr(人名) "、"ns(地名) "、"nt(机构名) "、"nz(其他专有名词) "、"vn(动名词) "、"an(名形词) "、"f(方位词) "等.

通过参考获得的决策表条件属性部分及其在原文中的上下文环境,人工进行相应的决策属性标注,然后只保留词性列(条件属性)和决策属性列作为粗集的输入进行规则获取.本文利用粗集工具 RSL[6]进行规则获取.为了确保规则的可靠性,采用集合下近似作为规则生成策略.

抽取出的规则具有如下的模式:

模式 1:$(p(w_L)==" * ")\wedge(p(w_1)==" * ")\wedge(p(w_2)==" * ")\wedge(p(w_R)==" * ")\Rightarrow k(w_1w_2)==" * "$;

模式 2:$(p(w_L)==" * ")\wedge(p(w_1)==" * ")\wedge(p(w_2)==" * ")\wedge(p(w_3)==" * ")\wedge(p(w_R)==" * ")\Rightarrow k(w_1w_2w_3)==" * "$.

模式 1 判断词性分别为 $p(w_1)$、$p(w_2)$ 的二个词构成的词序列在左边界词的词性为 $p(w_L)$,右边界词的词性为 $p(w_R)$ 的情况下,是否满足关键词短语的构成规则;模式 2 判断词性分别为 $p(w_1)$、$p(w_2)$、$p(w_3)$ 的三个词构成的词序列在左边界词的词性为 $p(w_L)$,右边界词的词性为 $p(w_R)$ 的情况下,是否满足关键词短语的构成规则.规则右边 $k(w_1w_2)$ 或 $k(w_1w_2w_3)$ 的取值为 1 或 0.

3 生成规则的后处理

不是所有符合规则的词序列都被最终抽取,还需要综合考虑词对文章主题的覆盖程度,即进行词的重要性评价.我们利用遗传算法通过线性加权的方法综合考虑各种相关特征进行文章中词的重要性评价.由于篇幅所限,这种方法的具体细

节将另文叙述. 对于重要性评价中权值较高的候选词,还需要进行冗余检测和消除. 相似度较大的两个词不适合同时被抽取,如"气象"和"天气"这样两个词. 本文利用知网[7]作为知识源来计算词的相似度并进行冗余检测和排除处理. 在本文的应用中,两个词的相似度算法如下:

算法 2　词的相似度算法

(1) 对于任意两个候选词 u、v;

(2) 到知网中检索这两个词,定义查找模式为精确匹配;

(3) 假设两个候选词的义项个数分别为 $S1$、$S2$;

(4) 对于候选词 u 的每个义项 $u(i)$,$0 \leq i \leq S1-1$;

(5) 对于候选词 v 的每个义项 $v(j)$,$0 \leq j \leq S2-1$;

计算 $u(j)$ 和 $v(j)$ 的概念相似度;

(6) 将义项相似度的最大值作为、的相似度.

词语相似度是一个主观性相当强的概念. 脱离具体的应用去谈论词语相似度,很难得到一个统一的定义[8]. 应该说,知网中任何两个不同词的所有义项之间的相似度有很大不同[9]. 但是如果这两个词出现在同一篇主题比较突出的文章中,则可以用相似度最大的义项之间的相似度代替两个词之间的相似度来解决冗余排除问题,这体现了相似度计算需要面向具体应用的观点,同时也是算法 2 的依据.

4　利用生成的规则进行关键词自动抽取

为了用有限数目的关键词表达尽可能多的信息,关键词抽取系统需要包括关键词短语的构成规则匹配、词的重要性评价和关键词冗余检测及消除等功能[10]. 本文采用如下步骤完成关键词的自动抽取:

(1) 输入待抽取关键词的文章;

(2) 进行分词、名实体识别、词性标注等处理,并过滤掉停用词;

(3) 保存文章的词向量以及对应的词性向量(两向量的元素个数相同);

(4) 统计每个词的出现词频和其他相关特征(包括词出现的特定位置、与某些线索词的同现信息等);

(5) 查找满足词性搭配规则的短语,并将其放入集合 C_1;

(6) 计算每个词的权重,并将权重最大的前 n_1 个词放入集合 C_2;

(7) 对集合 C_2 中的任何两个词进行评价,考察彼此是否存在字符串包含关系或者相似关系,并相应进行压缩;

(8) 考察 C_2 中的任何 $n(n=2,3)$ 个词的组合是否等于 C_1 中的某成员 P 或者是 P 的子字符串,如果是,则将短语 P 放到关键词集合 C_3 中;

(9) 如果 C_3 中的元素个数小于应该生成的关键词个数 n_2,则将候选词集合 C_2 中的其余词以权重大小为序放到 C_3 中,直到 C_3 中的元素个数等于 n_2 为止;

(10) 否则,只保留 C_3 中前 n_2 个短语.

其中,词的权重计算方法为相关特征的线性加权,参数采用遗传算法进行优化. 在实际应用中,取 $n_1=20$, $n_2=5$. 其中 n_1 为考察的候选词的个数,而 n_2 为最终生成关键词的个数. 系统优先抽取符合规则的关键词短语,并将最终的抽取结果保存在集合 C_3 中.

5　实验结果和分析

5.1　规则挖掘

本文人工标注了 7500 条中文关键词短语语料,用于挖掘关键词短语的构成规则. 语料中正例 4823 条,反例 2677 条. 语料取材于 1998 年 1 月份的人民日报电子版.

图 1 显示了训练语料规模和生成的规则数目的关系. 其中最上面的曲线表示至少需要 3 个正例支持,而最下面的曲线表示至少需要 12 个正例支持. 从图中可以看出,在训练语料规模相同的情况下,随着支持正例阈值的增加,规则的个数一般会随之减少. 每条曲线表示在支持正例阈值一定的情况下,挖掘出的规则数目与训练语料规模的关系. 开始时,规则个数随着训练语料的规模增加比较明显,在训练语料个数达到大约 5500 左右时,所有曲线开始趋于稳定. 特别是靠近下面的曲线比较明显,而靠近上面的曲线(正例阈值较小)则呈现微弱的增长趋势. 这是因为下面的曲线的阈值较大,支持正例阈值较大的规则个数是相对稳定的. 挖掘出的规则与训练语料规模有直接关系. 语料规模较少,则不足以挖掘出全部规则,挖掘出规则的可靠性也难以保证;只有达到一定的语料规模,才能获取全部规则,并保证其可靠性. 图 1 表明本文标注语料的规模可以满足规则挖掘的要求. 结合图 1 的结果,并通过对生成的规则进行人工校对,最终确定保留 94 条规则集成到系统中.

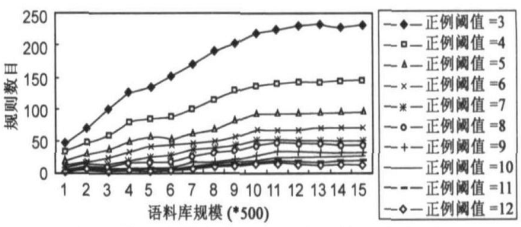


图 1　训练语料规模和生成规则数目的关系

5.2　规则评价

通过考察规则对关键词抽取系统性能的影响来对规则进行评价. 目前一种比较常用的关键词抽取性能评价方法是首先人工标注关键词,然后将系统抽取的关键词集合与人工标注的集合进行比对来进行评价. 这样的评价体系必须考虑语言表达方式多样性和标注者主观因素对评价结果的影响. 为此,本文采取多篇文档多个标注者评价的策略,目的是对系统的性能做出一个客观的评价. 由 5 个标注者对 100 篇新闻语料分别标注关键词,用于评价系统的性能.

假设共计有 N_e 个标注者标注文档,并且文章 i 中的关键词 p_{ij} 被 $N(p_{ij})$ 个标注者标注($0 \leq N(p_{ij}) \leq N_e$). 可以采用如下方法来评价系统,如果 p_{ij} 被超过 N_e'($0 \leq N_e' \leq N_e$)个标注者标注,即 $N(p_{ij}) \geq N_e'$,则可以认为系统在 p_{ij} 上的分值为 1,否则为 0. 这样系统的精确率定义为

$$P(N_e') = \frac{1}{N_t * N_k} \sum_{i=0}^{N_t-1} \left(\sum_{j=0}^{N_k-1} b_{ij}(N_e') \right) \quad (1)$$

如果 $N(p_{ij}) \geq N_e'$,则 $b_{ij}(N_e'=1)$,否则 $b_{ij}(N_e')=0$. 式(1)中 N_t 是测试集合中的文本个数. N_k 是被系统输出的关键

词个数,其默认值为 5.

系统的召回率定义为

$$R(N_e') = \frac{1}{N_t}\sum_{i=0}^{N_t-1}\left(\frac{1}{N_i}\sum_{j=0}^{N_i-1}e_{ij}(N_e')\right) \quad (2)$$

式(2)中 N_i 表示文本 i 中被超过 N_e' 个标注者标出的关键词的个数,如果其中一个关键词被系统抽取,则 $e_{ij}(N_e')=1$,否则 $e_{ij}(N_e')=0$.

精确率 $P(N_e')$ 表示系统抽取的关键词有多少(以及多大程度上)被标注者标出,召回率 $R(N_e')$ 表示系统是否将已经被超过一定数目的标注者标出的关键词抽取,二者的取值和阈值 N_e' 有关.

从图 2 可以看出,随着阈值 N_e' 的增加,召回率明显提高,而精确率有所下降. 这表明如果一个关键词被多数标注者标出,则系统一般可以将其抽取. 然而并不是系统抽取的所有关键词都会被多数标注者认可. 图 3 考察了在 $N_e'=2$ 时,施加不同的规则个数对系统性能的影响. 在不施加任何规则的情况下,系统的性能最低. 这表明单纯的抽取词作为关键词对原文主题的覆盖能力是较低的. 随着规则的增加,系统的性能明显提高. 当规则的个数增加到一定数值时,系统的性能趋于稳定.

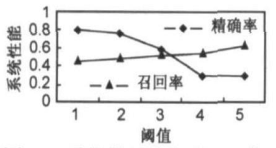


图 2 系统性能与阈值 N_e' 的关系

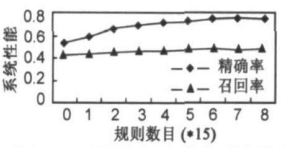


图 3 系统性能与施加的规则个数的关系 ($N_e'=2$)

同时将本文方法与传统通过 N-gram 频率统计方法挖掘的规则进行了比较. N-gram 频率统计方法是通过统计构成关键词短语的词的词性组合频率,来发现关键词短语的词性匹配模式. 通过引入 N-gram 方法获取的规则,系统的精确率为 0.69,召回率为 0.45,均低于引入全部粗集规则后系统的相应指标(精确率为 0.76,召回率为 0.48,如图 3 所示).

6 结论

本文利用粗集理论对关键词短语的构成规则进行了挖掘,结果表明粗集的知识约简和规则发现能力比较适合进行关键词短语构成规则的挖掘工作. 将挖掘出的规则用于指导关键词的自动抽取,避免了一些错误的搭配被抽取,从而提高了系统的性能,使抽取结果更加符合人们的习惯. 本文的另外一个贡献是提出了一种评价体系,其目的是为了客观评价关键词抽取系统的性能.

参考文献:

[1] 李素建,王厚峰,余士汶,辛乘胜. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报,2004,27(9):1192 - 1197.
Li Su-Jian, Wang Hour-Feng, Yu Shi-Wen, Xin Cheng-Sheng. Research on maximum entropy model for keyword indexing [J]. Chinese Journal of Computers, 2004, 27(9): 1192 - 1197. (in Chinese)

[2] 韩客松,王永成. 全文标引的主题词标引和主题概念标引方法[J]. 情报学报,2001,20(2):212 - 216.
Han Ke-song, Wang Yong-cheng. Methods of keyword and subject concept indexing to chinese full-text [J]. Journal of the China Society for Scientific and Technical Information, 2001, 20(2): 212 - 216. (in Chinese)

[3] Anette hulth. Combining machine learning and natural language processing for automatic keyword extraction [D]. Stockholm: Department of computer and systems sciences, Stockholm University, 2004. 35 - 38.

[4] 俞士汶,陆俭明,朱学锋,等. 现代汉语语料库加工规范——词语切分与词性标注 [S]. http://www.icl.pku.edu.cn/icl_groups/corpus/spec.htm, 1999.

[5] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341 - 356.

[6] Michal, Jacek. Rough set theory library. http://www.pw.edu.pl/english/, 1994.

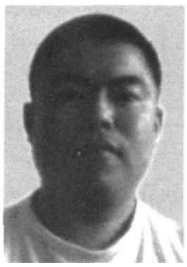
[7] 董振东,董强. 知网. http://www.keenage.com/, 2004.

[8] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[A]. 第三届汉语词汇语义学研讨会论文集[C]. 台北, 2002. 59 - 76.

[9] 卢志茂,刘挺,李生. 统计词义消歧的研究进展[J]. 电子学报,2006,34(2):333 - 343.
Lu Zhi-mao, Liu Ting, Li Sheng. The research progress of statistical word sense disambiguation [J]. Acta Electronica Sinica, 2006, 34(2): 333 - 343. (in Chinese)

[10] 王晓龙,关毅,等. 计算机自然语言处理[M]. 北京:清华大学出版社,2005. 128 - 129.

作者简介:



刘远超 男,1971 年生于黑龙江双城,哈尔滨工业大学计算机学院在职博士,讲师,1995 年毕业于哈尔滨工业大学,获工学学士学位,1997 年毕业于哈尔滨工业大学,获工学硕士学位,主要研究方向:自然语言处理、人工智能.
E-mail: lyc@insun.hit.edu.cn