基于词语热度的启发式中文句子压缩算法

韩 静,张东站
HAN Jing, ZHANG Dongzhan

厦门大学 信息科学与技术学院,福建 厦门 361005
School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China

HAN Jing, ZHANG Dongzhan. Heuristic Chinese sentence compression algorithm based on hot word. Computer Engineering and Applications, 2014, 50(4):132-139.

Abstract: Since the parallel sentence/compression corpora which most of the traditional methods based on are not easy to obtain, a linguistically-motivated heuristics Chinese sentence compression algorithm is proposed after studying traditional methods. By analyzing the human-produced compression and linguistic knowledge, two sets of rules are proposed, one is in word layer and the other is in clause layer. Two sets of rules based on the parse tree and the words dependence are used to compress sentence, and enhance the algorithm by hot word in order to keep the algorithm flexibility and accuracy. In the last step the compression result is cleaned and repaired. Human-produced compression, rule-only algorithm and hot word enhanced algorithm are compared then the results are evaluated in compression rate, grammaticality, informativeness and heat. The experimental results show that heuristic Chinese sentence compression algorithm based on hot word can improve the heat of compression results without much loss in compression rate, grammaticality and informativeness.
Key words: Chinese sentence compression; hot word; linguistic; parse tree

摘 要: 传统的句子压缩方法多基于难以获得的“原句-压缩句”对齐语料库,因此提出了不依赖于对齐语料库的中文句子压缩算法。通过研究人工压缩结果并结合语言学知识,提出了词语层面和分句层面的两组压缩规则。算法在原句句法分析树和词语间依赖关系的基础上,使用两组规则进行压缩,同时为了保证压缩算法具有更强的适应性和准确性,引入词语的热度加强了压缩算法,最后通过句子整理和语法修复得到最终的压缩句。对比了人工压缩、只使用规则压缩和引入词语热度压缩三种压缩方法。实验结果表明,基于热度的启发式中文句子压缩算法可以在压缩比、语法性、信息量都损失较少的情况下,提高压缩句的热度。
关键词: 中文句子压缩;热词;语言学;句法分析树
文献标志码: A **中图分类号:** TP391.1 doi:10.3778/j.issn.1002-8331.1305-0261

1 引言

随着网络规模的迅速扩大和接入网络的日趋便利,人们从网络获取信息的数量也飞速增长。在获取信息数量提升的同时,人们也希望精简信息以节约时间。近年来自然语言处理的飞速发展,使计算机渐渐参与到这个工作中来,句子压缩技术就是其中一项重要技术。句子压缩是最重要的文本重写技术之一,其主要目的是在保持语法规范的前提下,提取出句子的主要信息。句子压缩的应用非常广泛,并已经在自动摘要/文本压缩[1]、电子邮件提醒[2]、聚合内容(Rich Site Summary,RSS)[3]以及移动终端[4]领域等诸多方面投入使用。但是传统方法自动生成的压缩句,不能跟上互联网信息热点转移的步伐,为了满足用户最新最热信息的需求,本文算法中考虑了词语的热度,更多采用用户关注的热门词语,保留更热的信息。

句子压缩就是把给定的原始的长句子改写成一个短句子,这个短句子要保留原句中的重要信息,并且符合语法规范。句子压缩因为用途广泛,近年来受到了很大的关注,尤其是自动摘要和自动标题。一个理想的句子压缩算法包含复杂的改写操作,比如词语的替换、插

基金项目: 国家自然科学基金(No.50604012)。
作者简介: 韩静(1987—),女,硕士研究生,研究领域为自然语言处理;张东站,男,博士,副教授,研究领域为数据库理论与应用、自然语言处理。E-mail:zdz@xmu.edu.cn
收稿日期: 2013-05-21 **修回日期:** 2013-09-30 **文章编号:** 1002-8331(2014)04-0132-08

入、重排序等。本文只关注简单的词语删除——删除原句中次要的词语。

本文给出了一个不依赖“原句-压缩句”对齐语料库训练的中文句子压缩算法,将基于语言学的启发式规则和词语的热度相结合,不断降低词语在句子中的权重,最终将权重高于阈值的词语组合成为压缩句。本文算法只考虑词语的词性、词语间的依赖关系和句子的句法分析树,降低了生成启发式规则的难度和语言学技巧。

2　相关工作

2.1　国内外研究进展

目前多数的句子压缩算法是基于统计学习的方法,其中又以依赖“原句-压缩句”对齐语料库的算法居多,将对齐语料库作为训练语料库,学习原句与压缩句之间的对应关系。通常从句子的句法分析树中提取特征,用来学习原句到压缩句之间的转换规则或者评估候选压缩句的得分。

目前针对英文的句子压缩算法比较丰富,Knight,Marcu(K&M)[5]提出了采用机器学习进行句子修剪的方法,分别采用了基于噪声信道模型(noisy-channel model)的方法和基于决策树的模型,为后来许多研究奠定了基础。Nguyen等[6]提出了基于支持向量机的模型,McDonald[7]提出了最大边缘方法,使用这些判别模型可以降低机器学习的错误率。

目前大多数句子压缩算法普遍采用“原句-压缩句”的对齐语料库,不使用或者少使用对齐语料库的算法却很少。由于“原句-压缩句”的对齐语料库目前很少且很难获得,当前大多数算法的实际应用价值不高,因此实际应用中急需研究一种能够脱离“原句-压缩句”的对齐语料库的压缩算法。Hori和Furui[8]采用一种无监督的方法删除句子中的词语,利用动态规划算法优化句子修剪过程,但他们的方法并没有考虑句法方面的信息。Turner和Chamiak[9]改进了K&M的方法,并分别采用有监督、无监督、半监督三种方法下的噪声信道模型。Clarke和Lapata[10]采用整数规划算法对句子压缩过程进行全局优化。

相对于丰富的英文句子研究成果,中文句子压缩研究很少。Wei Xu和Ralph Grishman[11]在语言学启发式规则的基础上,通过事件词语意义(event word significance)和事件信息密度(event information density)来加强算法对词性标注和句法分析的容错性能。赵青[12]给出了一个基于概率统计的句法分析的中文句子压缩系统,该系统引入了有监督的机器学习方法来提取压缩规则,通过统计原句和压缩句在压缩前后句法成分的变化规律来计算各个句法成分的删除概率。

本文算法与Dorr等[13]的自动标题算法和Wei Xu和Ralph Grishman的算法最为相似,都是使用语言学的启发式规则来修剪句法分析树。这种方法按照规则不断删除句法分析树中的次要成分,直到达到指定的压缩句长度。这种方法通过按照风险由低到高的顺序应用规则来降低风险,保证压缩句的语法性。

2.2　WeiXu和Ralph Grishman算法介绍

以下给出WeiXu和Ralph Grishman的启发式规则、事件词语意义及事件信息密度。

Set 0-basic:

(1)括号中的成分;

(2)副词(否定副词、一些时序副词、程度副词除外);

(3)形容词(被修饰的名词只有一个词语时除外);

(4)DNPs(以多种类别短语组合以“的”结尾,并且作为NP的修饰成分出现);

(5)DVPs(以多种类别短语组合以“地”结尾,作为VP的修饰成分出现);

(6)名词并列短语中除了第一个名词以外的词语。

Set 1-fixed:

(1)NP节点的所有子节点,除了时序词、专有名词和最后一个名词词语;

(2)如果句子包含多个简单句(IP),除了第一个简单句的所有简单句;

(3)介词词组,除了包含位置信息和时间信息(根据人工收集的一些介词)。

Set 2-flexible:

(1)动词并列短语中除第一个外的所有节点;

(2)关系从句;

(3)同位语从句;

(4)所有介词短语;

(5)NP节点中除第一个节点外的所有子节点;

(6)如果句子中包含多个IP,句子中的所有简单句。

以上启发式规则只是用来检测哪些节点是可删除的,真正的删除操作在一个考虑词语意义和压缩率的压缩句产生和选择模块中进行。

(1)词语意义

由于以上规则压缩过程中会优先去掉修饰成分,即使这些修饰词很重要,因此WeiXu等人提出了词语意义来评价词语的重要程度。

人工压缩的句子主要描述了事件或者一系列相关的事件并包含很大比例的名称实体(尤其在新闻领域中),与基于事件的摘要[14]相似,他们只考虑事件词语,即名词、动词和专有名词。

词语意义描述了词语 w_i 对文章 j 的重要性:

$$I_j(w_i)\begin{cases} tf_{ij}\times idf_i, \text{ if } w_i\text{是动词或者普通名词} \\ tf_{ij}\times idf_i+\omega, \text{ if } w_i\text{是专有名词} \\ 0, \text{其他} \end{cases}$$

其中,w_i 为文章 j 句子中的词语,tfi_j 为文章 j 中 w_i 的频率,idf_i 为 w_i 的逆文档频率,w 为专有名词的附加

权重。

(2)压缩句生成及选择

此模块中采用贪心算法修剪已经经过标注权重的句法分析树,修剪算法如下:

①删除一个权重最低的节点,得到一个候选压缩句;

②更新删除节点的所有祖先节点的权重;

③重复上两个步骤直到没有节点可以删除。

选择的过程是一个句子长度和信息量权衡的过程,该算法提出了信息密度来衡量句子的信息量,信息密度公式如下:

$$D(s_k)=\frac{\sum_{w_i\in P} I(w_i)}{L(s_k)}$$

该算法最后将信息密度最高的句子作为系统的压缩句输出。

3　基于热度的启发式中文句子压缩算法

3.1　算法总体流程

本文算法的总体流程如图1所示,首先使用ICTCLAS将句子分词,将去掉词性标注的分词结果作为Stanford Parser句法分析器的输入,输出的句法分析树和词语之间的依赖关系作为本文压缩算法的输入,在此基础上做句法分析树修剪操作,从而生成初步压缩句,生成的压缩结果再通过句子整理和语法修复两个模块处理,生成最终的压缩句。其中压缩算法有两种,一种是只基于语言学的启发式规则(Rule算法),另一种是在启发式规则的基础上同时考虑词语的热度(Hot算法)。

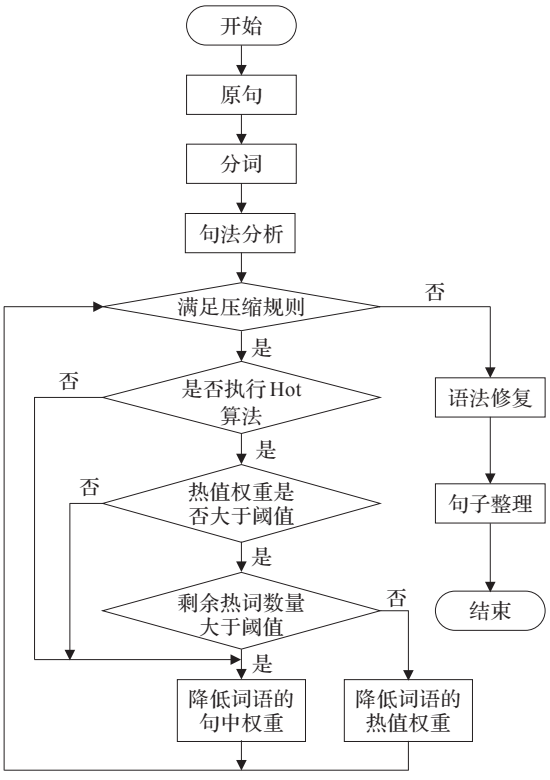

图1　系统总体流程图

包括压缩算法在内的所有模块将在下面各小节中详细介绍。

3.2　预处理

3.2.1　中文分词

分词是将一个完整的句子划分成词语序列的过程,文本预处理的首要任务就是分词,分词的关键在于如何选择恰当的语义单元。在英文中,以空格作为单词间的自然分界符,可以方便地将文本转化为语义单位组成的词语序列。

与英文相比,中文分词的难度更高,因为中文文本中没有形式上类似于空格的标示词边界的分隔符。目前常见的分词方法有三种:基于字符串匹配的方法、基于理解的分词方法和基于统计的方法。到底哪种分词算法的准确度更高,目前尚无定论。不过分词系统一般采用多种方法结合的方式分词,不依赖单一的算法来实现,以取得更好的分词效果。

本文采用中科院研发的ICTCLAS分词软件,该软件分词速度快,准确率高。

“我爱北京天安门”分词和去掉分词性标注结果如下例所示。

例:我/r 爱/v 北京/ns 天安门/ns。/w 我 爱 北京 天安门。

3.2.2　句法分析

句法分析主要是分析句子的结构方式,考察句子结构内部各组成成分的特点,描写句子的结构规则。句法分析的任务是自动分析出句子的语法结构及语法关系,将一个线性序列的句子转换成一个结构化的语法树。

图2为句子“我爱北京天安门”的句法分析树和依赖关系,左侧的句法分析树中ROOT表示树的根节点,每组匹配的括号内为句子的一种成分;右侧的依赖关系表示出句子中后一个词语对前一个词语的依赖性质。

```
(ROOT
  (IP
    (NP(PN我))
    (VP(VP爱))
      (NP(NR北京)(NR天安门))
    (PU.)))
```

(a)句法分析树

```
nsubj(爱-2,我-1)
root(ROOT-0,爱-2)
nn(天安门-4,北京-3)
dobj(爱-2,天安门-4)
```

(b)依赖关系

图2　句法分析树与依赖关系

句法分析树和依赖关系的使用有助于判断词语在句子中的重要程度。本文采用美国斯坦福大学自然语言处理小组开发的Stanford Parser作为句法分析器,语言学模型使用适用简体中文且分析速度较快的xinhua-PCFG模型。

3.3　基于语言学的启发式规则

启发式算法是相对于最优算法提出的。一个问题的最优算法求得该问题每个实例的最优解。启发式算法可以这样定义:一个基于直观或经验构造的算法,在

可接受的花费(指计算时间和空间)下给出待解决组合优化问题每一个实例的一个可行解,该可行解与最优解的偏离程度不一定事先可以预计[15]。

由于使用规则描述自然语言现象具有长尾效应,即描述所有语言现象的完整规则难以实现,因此本文采用启发式规则。启发式规则试图用数量较少的规则,提供执行效率良好且可得到相对满意的压缩句的规则。启发式规则处理句子压缩问题时通常可以在合理时间内得到不错的压缩句。

基于语言学的启发式规则旨在识别句法分析树中可以去掉的成分,并且尽量少损失语法性和重要信息。基于对人工压缩的句子的研究,并结合语言学知识,本文得出以下启发式规则,这些规则按照去掉后对句子的影响排列,并分成了两个集合:

Set_0—词语层面压缩规则:

定义1　S 是一个句子,则 S 可以表示成 $S=t_1t_2\cdots t_i\cdots t_j\cdots t_n$ 的形式,其中 S 中的词语或标点符号称为token,用 t_i 表示。

(1)在句子 $S=t_1t_2\cdots t_i\cdots t_j\cdots t_n$ 中,Set_{B} 为 S 中小括号或中括号中词语的集合(包括括号本身),如果 $t_i\cdots t_j\in Set_{\mathrm{B}}$,则 $t_i\cdots t_j$ 可以去掉。

小括号和中括号中的成分一般起到解释说明的作用,去掉之后对句子整体影响很小。

(2)在句子 $S=t_1t_2\cdots t_{i-1}t_i\cdots t_j\cdots t_n$ 中,Set_{NP} 为 S 中名词短语(NP)构成的集合,如果 $t_i\cdots t_j\in Set_{\mathrm{NP}}$ 且 $t_1t_2\cdots t_{i-1}\notin Set_{\mathrm{NP}}$,则 $t_1t_2\cdots t_{i-1}$ 可以去掉。

句子的第一个NP,一般是句子的主语,而第一个NP之前的成分通常是句子的状语,状语是修饰限制中心语的成分,去掉对句子的主要意思影响较小。

(3)句子 $S=t_1t_2\cdots t_{i-1}t_i\cdots t_j\cdots t_n$ 中,Set_{ADV} 为 S 中副词(否定副词和疑问副词除外)构成的集合,如果 $t_i\cdots t_j\in Set_{\mathrm{ADV}}$,则 $t_i\cdots t_j$ 可以去掉。

副词用来修饰限定形容词和动词,表示否定含义的副词如果去掉会导致句子的意思相反,因此不能去掉,疑问副词(如"怎样"等)一旦去掉对会使句子由疑问句变为陈述句,因此也不能去掉。

(4)句子 $S=t_1t_2\cdots t_i\cdots t_jt_{j+1}t_k\cdots t_n$ 中,Set_{ADJ} 为 S 中形容词构成的集合,如果 $t_i\cdots t_j\in Set_{\mathrm{ADJ}}$ 且 $t_{j+1}\cdots t_k\in Set_{\mathrm{NP}}$ 则 $t_i\cdots t_j$ 可以去掉。

修饰限定名词的形容词短语在句子中做定语,属于次要成分,一般来说可以删掉。

(5)句子 $S=t_1t_2\cdots t_i\cdots t_j\cdots t_n$ 中,Set_{PQ} 为 S 中作定语的介词短语(PP)或者数量词短语(QP)构成的集合,$t_i\cdots t_j\in Set_{\mathrm{PQ}}$,则 $t_i\cdots t_j$ 可以去掉。

PP做定语一般来说比较冗长,而且去掉后句子中心意思影响较小,QP做定语一般表示名词数量做定语,不涉及到句子的中心意思,因此两者都可以去掉。

(6)句子 $S=t_1t_2\cdots t_i\cdots t_jt_{j+1}t_k\cdots t_n$ 中,Set_{DNP} 是修饰NP的以"的"结尾的短语(DNP),如果 $t_i\cdots t_j\in Set_{\mathrm{DNP}}$ 且 $t_{j+1}\cdots t_k\in Set_{\mathrm{NP}}$,则 $t_i\cdots t_j$ 可以去掉。

(7)句子 $S=t_1t_2\cdots t_i\cdots t_jt_{j+1}t_k\cdots t_n$ 中,Set_{DVP} 是修饰动词短语(VP)的以"地"结尾的短语(DVP)构成的集合,如果 $t_i\cdots t_j\in Set_{\mathrm{DVP}}$ 且 $t_{j+1}\cdots t_k\in Set_{\mathrm{VP}}$,则 $t_i\cdots t_j$ 可以去掉。

DNP和DVP是修饰名词短语和动词短语的短语,即在句子中做定语,属于次要成分,因此可以去掉。

(8)句子 $S=t_1t_2\cdots t_i\cdots t_j\cdots t_n$ 中,Set_{PP} 为 S 中作状语的PP构成的集合,如果 $t_i\cdots t_j\in Set_{\mathrm{PP}}$,则 $t_i\cdots t_j$ 可以去掉。

介词短语做状语一般来说比较冗长,而且去掉后句子中心意思影响较小。

Set_1—分句层面压缩规则:

定义2　句子 S 有 $S_1,S_2,\cdots,S_n$ 共 n 个分句,则 S 可以表示成 $S=S_1,S_2,\cdots,S_n$。

定义3　如果句子 S 是主从关系复合句,则 S 可以表示为 $S=S_1,S_2,\cdots,S_i\Rightarrow S_{i+1},\cdots,S_n$,其中 $\Rightarrow$ 表主从关系,S_{i+1} 为主,S_i 为从。

定义4　如果句子 S 是并列关系复合句,则 S 可以表示为:

$$S=S_1,S_2,\cdots,S_i\vee S_{i+1}\vee S_{i+2}\vee\cdots\vee S_{i+j},S_{i+j+1},\cdots,S_n$$

其中 $\vee$ 表并列关系。

(1)如果 $S=S_1,S_2,\cdots,S_i\Rightarrow S_{i+1},\cdots,S_n$,则 S_i 可以去掉。

主从关系复合句一般由转折、递进、因果等关系,转折中常见的标志有"虽然"、"但是"等,递进中常见的标志有"不但"、"而且"等,因果中常见的标志有"因为"、"所以"等。

(2)如果 $S=S_1,S_2,\cdots,S_i\vee S_{i+1}\vee S_{i+2}\vee\cdots\vee S_{i+j},S_{i+j+1},\cdots,S_n$,则 $S_{i+1},S_{i+2},\cdots,S_{i+j}$ 可以去掉。

如果句子包含多个并列分句,可以适当去掉后面的并列分句。

(3)如果 $S=S_1,S_2,\cdots,S_n$,则 $S_2,S_3,\cdots,S_n$ 可以去掉。

一般重要的分句在句子中的位置靠前,因此选择保留第一个分句。此规则对比前面所有规则风险最大,因此最后使用。

Set_0 是在词语层面压缩,并不涉及句子的整体结构,对句子的语法性与信息量影响较小;而 Set_1 中的规则在分句层面进行压缩,由于一次去掉一个分句,会对句子的结构产生影响,因此本文算法按照 Set_0、Set_1 的顺序应用压缩规则,以减少压缩过程中信息量的损失。

Rule算法如下所示,其中 *SenTokens* 为 S 的去掉标

注的token列表,satisfy函数根据 S 的句法分析树和依赖关系集合计算 *SenTokens* 中满足规则rule的词语集合,并返回该集合。每个token对应一个句子中权重 *senWeight*, *senWeight* 初值为0,对每个token依次应用 Set_0 和 Set_1 中的规则,如果满足该rule的token,减少其 *senWeight*,Rule算法中每次减少步长为5。如果希望压缩句的语法性较好且对压缩句长度没有要求,此时可以结束循环,将所有 *senWeight* 非负的token组成压缩句。如果希望获得更短的压缩句,还可以继续应用 Set_1 中的规则进行压缩,直到压缩到满意的长度为止。Rule算法虽然将 Set_0 和 Set_1 写到一起处理,但是实际上在 Set_1 中任意rule执行后均可退出循环。

算法 Rule Algorithm

输入:S 去掉词性标注的token列表 *SenTokens*;S 的句法分析树 T 和依赖关系集合 D

输出:S 的压缩句 C

方法:

(1)Initialize senWeight

(2)for each rule in Set_0 and Set_1 //可以在执行完任意 Set_i 中rule后退出循环

(3) Set=satisfy(SenTokens,rule, T, D)

(4) for each t in Set

(5) reduce t's senWeight

(6) end for

(7)end for

(8)C consists of tokens with nonnegative senWight not

3.4 基于热度的压缩方法

3.4.1 热词及热度

热词即热门词汇、热搜词,热词具有时代特征,其作为一种词汇现象,反映了一个时期内人们普遍关注的热点话题。词语的热度,即某个词语在过去一段时间内的网络曝光率以及用户的关注度和媒体的关注度,它能够反映人们关注的程度。将句子压缩与热词结合起来可以增加压缩结果的热度,紧跟时代趋势,保留更多的热词,使之能够更容易被用户搜索到,因此将句子压缩与热词结合起来是有必要的。

目前词语热度主要有谷歌趋势和百度指数两种获取方式。谷歌趋势是通过全球搜索结果分析得来的,百度主要是对应中文用户。鉴于百度指数结果更容易获取且是更针对中文的分析结果,因此本文采用百度指数计算词语的热值。

3.4.2 基于热度的压缩方法

WeiXu等人算法提到使用tf-idf来计算词语意义,tf-idf的文章范围为实验数据集中的文章,本文认为在这样的限定范围内计算词语的重要性不可取,因此引入更大范围的基于热门搜索引擎统计的词语热度来表示词语重要性。

Hot算法如下所示,词语的热值hotWeight按照公式(1)初始化,标点的 *hotWeight* 初始化为0。其中 I 为词语的百度指数,百度指数取值区间很大,取对数之后即可以减少各个词语之间热度的差距,又能保证趋势不变。有些词语的百度指数为0,因此将 I 加1可保证真数大于1,取对数之后得到的 *hotWeight* 非负。*hasChange* 表示一轮循环中所有的 *senWeight* 和 *hotWeight* 是否有变化,*hasChange* 的初值为false。

$$H=\lg(I+1) \tag{1}$$

算法 Hot Algorithm

输入:S 的token列表 *SenTokens*,S 的句法分析树 T 和依赖关系集合 D,至少保留热词数量remainThreshold

输出:S 的压缩句 C

方法:

```
(1)Initialize senWeight and hotWeight
(2)do
(3)  hasChange =false
(4)  for each rule in Set0
(5)    Set=satisfy(SenTokens,rule,T,D)
(6)    for each t in Set
(7)      if t's hotWeight>0
(8)        if the |HotWords|>remainThreshold
(9)          reduce the hotWeight of t
(10)         hasChange=true
(11)         if t's hotWeight<=0
(12)           delete t from HotWords
(13)         end if
(14)       end if
(15)     else
(16)       reduce t's senWeight
(17)       hasChange=true
(18)     end if
(19)   end for
(20) end for
(21)while hasChange=true
(22)  if a shorter C is needed
(23)    for each rule in Set2
(24)      Set=satisfy(SenTokens,rule,T,D)
(25)      for each t in Set
(26)        reduce t's senWeight
(27)      end for
(28)    end for
(29)  end if
(30)  C consists of tokens with nonnegative senWight not
```

首先对句子循环应用 Set_0 中的规则,如果满足压缩规则的词语是热词,并且句子中剩余的热词数量小于 *remainThreshold*,则减少该热词的 *hotWeight*,而不减少 *senWeight*,*hasChange* 置为true;如果该词语不是热词或者

剩余热词的数量不小于*remainThreshold*，则直接减少该词语的*senWeight*，*hasChange*置为true。如果*hasChange*置为true，则继续循环应用 Set_0 中的规则，否则表示一轮循环过后压缩规则没有任何效果，则退出循环。由于应用规则的效果会不断叠加，因此循环使用 Set_0 会使满足多个规则的和热度较低的词语会具有较低的*senWeight*，有助于保留更热更重要的词语。当剩下的成分都比较重要（本文中为句子的主语、谓语和宾语）或者到达删除热词的阈值（本文中将阈值设为0）时停止 Set_0 的循环执行，如果对压缩结果的长度没有要求，此时可以结束算法，如果希望获得更短的压缩句，还可以继续应用 Set_1 中的规则进行压缩，直到压缩到满意的长度为止。Hot算法中*senWeight*和*hotWeight*每次减少步长均为5。

3.5　句子整理及成分修复

3.5.1　句子整理

经过以上各规则的处理，已经可以得到较为满意的压缩句了，但是因为压缩大多在词语层面进行，难免出现分句字数过少，或者标点符号缺失或过多的情况，通过句子整理模块可以删除字符数过小的分句，例如分句字数小于3个，则去掉该分句，删除多余的标点（如果一个分句中所有词语均被删除，会留下多余的逗号），对标点进行更正等操作（如句子的最后一个分句及句号被删除，句子以逗号结尾或者没有标点符号）。句子整理模块功能简单，但是不可或缺。

3.5.2　语法修复

句子中有些成分是相关联的，但是由于这些成分间的热度不同以及在句子中的重要程度不同，导致压缩后相关联成分被部分删除，句子不通顺，因此在算法的最后进行语法修复，以保证句子的语法性。主要分为以下几种情况：

（1）数词与量词：数词与量词是成对出现的，如果一个成分在压缩句中保留，则添加另一个成分到压缩句中。

（2）动词开头的压缩句：会出现这个问题是因为有些句子的主语会单独以短句的形式给出以示强调，但是由于分句过短会被句子整理模块去掉，因此导致压缩句以动词开头。如果压缩句以动词开头，则添加该动词之前的最近的依赖于该动词的NP到压缩句中。

（3）“的”字开头的压缩句：将“的”字去掉。

（4）书名号的处理：书名号中的内容一般是书名或者影视剧名称，应该全部保留或删除，由于在分词过程中会将书名号中的短语进行分词，因此需要在压缩完成后对书名号做特殊处理（为了降低压缩算法的复杂性，因此在压缩的过程中不考虑书名号的特殊性）。如果压缩句中有书名号中的词语，则把整个书名号中的内容全部添加到压缩句中，包括书名号本身，否则去掉书名号。

4　实验结果及分析

4.1　实验设置

由于目前中文句子压缩算法研究较少，目前尚未有公开的测试集，因此本文使用独立收集的测试数据评价算法。本文的测试数据来全部自于新浪网站的新闻，选自体育、科技、财经、娱乐等多个领域，类别范围广，具有代表性。本文选取了上述范围内的2 126条句子作为实验的数据集。

词语的句中权重初始值全部为0，热词的初始热值为由公式（1）计算出的热值，如果不是热词，则热值为0。在压缩的过程中，对于满足压缩规则的词语，如果该词语不是热词，就降低该词语的句中权重，如果是热词，就先降低其热值，当热值降到0时，就把该词语当作普通词语处理了。算法执行结束之后，将句中权重为负的词语组合起来，组成压缩句。

人工压缩[Manual]：人工地将一个完整的句子压缩成为一个较短的完整的句子，只能使用删词的方法，不能改变词语的顺序。

不考虑热度只是用规则的压缩[Rule]：即Rule算法，不考虑词语的热度，只是用上述规则进行压缩。

考虑热度的压缩[Hot]：即Hot算法，输入句子中热值最高的5个词语，压缩结果一般保留其中的3个。

4.2　实验结果评价

本文的句子压缩算法采用人工评价语法性和信息量，采用机器评价压缩比和热度比。人工评价的得分难免会受主观影响，因此本文制定了评价标准，可以降低人工评价的误差。以下是人工评价及机器评价标准：

（1）压缩比：压缩句的字符数/原句的字符数。

（2）语法性：分为1~5分5个等级，5分最好，1分最差，语法性的评价只与是否符合语法有关，与原来句子是否意思相同无关。

（3）信息量：信息量的评价都是把分句最重要的信息表达出来即可（不考虑语法性，不需要通顺，重要词语或短语出现即可），不需要完整成分。

（4）热度比：词语的热值为 H，句子的热度为句中所有词语热值累加所得。压缩句与原句词语热度通过求比值得到归一化的数值，热度比越接近1表示保留原句中热词比率越高。

$$CompHeatRate=\frac{\sum_{y_j\in Y}H_{y_j}}{\sum_{x_i\in X}H_{x_i}} \tag{2}$$

表1中显示了人工、WeiXu等人方法、Rule算法和Hot算法四种压缩方法的评价结果，从表中看出，人工压缩的句子压缩比最好，语法性也最好，但是由于没有考虑热词，热度比相对较低，WeiXu等人的算法压缩比不错，语法性略高与本文的两种算法，但是因为压缩规则

中词语等级与分句等级的压缩规则交替出现,因此包含的信息量较少,热度方面均比本文的两种算法差一些。Rule算法的压缩比和语法性都略好于Hot算法的压缩结果,但是在热度比方面有差距,热度提高得到的好处高于压缩比、语法性降低的损失。表中也可以看出两个结果算法差距并没有很明显。四个压缩方法的压缩比大致介于0.55到0.6之间,说明并没有压缩掉很多词语,因为测试数据全部来自于新闻,新闻的特点是用词简洁,修饰成分相对较少,所以压缩比较高。语法性方面自动压缩算法比人工压缩算法的差距还是很明显的,一些规则的应用是有风险的,并不适用所有的情况,在压缩掉一些句子中次要成分的同时也删除了另外一些句子中的重要成分。信息量方面人工压缩的结果信息量为4.613 05,基本上保留了原句中的重要信息,而两个自动压缩算法的信息量分别为4.123 86和4.202 37,相差无几,相对人工压缩的结果差了一些。本文两个自动压缩算法的信息量如此接近,说明多保留一些热词对压缩句信息量的影响微乎其微。

表1 评价结果

算法	压缩比	语法性	信息量	热度比
Manual	0.565 659	4.867 32	4.613 05	0.356 938
WeiXu	0.578 466	4.067 57	3.898 66	0.445 752
Rule Algorithm	0.589 555	3.810 53	4.123 86	0.533 865
Hot Algorithm	0.610 941	3.778 95	4.202 37	0.571 014

4.3 典型示例

例1~3列举出几个具有代表性的压缩结果示例。

下面为对各示例的分析说明。

例1

[O]但是,印度的一家政府机构最近建议政府禁止中兴通讯和华为参与该国国有光纤网络项目的竞标。

[M]印度的政府机构建议政府禁止中兴通讯和华为参与竞标。

[R]印度的政府机构建议政府禁止中兴通讯和华为参与竞标。

[H]印度的政府机构建议政府禁止中兴通讯和华为参与竞标。

例2

[O]在中国开始改革开放时,侯为贵于1985年前往深圳创办了一家电信设备公司,这家公司后来成为中兴通讯。

[M]侯为贵创办了中兴通讯。

[R]侯为贵创办了一家电信设备公司,这家公司后来成为中兴通讯。

[H]侯为贵前往深圳创办了一家电信设备公司,这家公司后来成为中兴通讯。

例3

[O]中兴公司今年第三季度亏损3.1亿美元,成为自该公司上市以来第一个出现亏损的季度。

[M]中兴公司今年第三季度亏损,成为第一个亏损的季度。

[R]中兴公司今年第三季度亏损3.1亿美元,成为季度。

[H]中兴公司今年第三季度亏损3.1亿美元,成为季度。

例1中2个自动压缩结果与人工压缩结果一样,在保证了语法通顺的同时又保留了原句中的主要信息,压缩结果令人满意。

例2展示了考虑热词后压缩句的热度的提升效果,Rule算法压缩句热度比为0.356 380,Hot算法压缩句热度比为0.477 422,可见热度有显著提升,主要差距在于“深圳”这一热词,“深圳”一词的热度在原句中仅次于“电信”,Hot算法会将“深圳”保留下来以加强压缩句热度。

例3展示了由于启发式规则不通用导致的错误,“自该公司上市以来第一个出现亏损的”符合Set 0中的第六条规则,符合这个规则的短语一般是名词的定语,这个句子中也是作为“季度”的定语出现,但是却不能按照此规则删除。

4.4 错误分析

造成压缩句出错的主要原因有以下几点:

(1)分词出错;

(2)句法分析树出错;

(3)启发式规则应用出错;

(4)考虑热词导致出错。

其中分词错误及句法分析树错误的例子本文没有给出,因为这两部分不是本文主要研究范围,因此只做错误分析。目前分词和句法分析树的正确率虽然都有改善,但是也达不到100%,因此使用其输出的结果必然会对最后压缩的效果造成影响。

启发式规则造成压缩句出错如示例3所示。由于使用规则描述自然语言现象具有长尾效应,即使使用再多的规则也无法完整描述自然语言现象,难免产生规则冲突、错误。

热词错误说明在压缩句中引入热词可以提高句子热度的同时也会增大压缩句出错的概率,但是鉴于单纯考虑热词导致压缩句错误的比例很小,引入热词还是值得的。

5 结论及展望

本文提出了一种新颖的中文句子压缩方法,该方法将基于语言学的启发式规则与词语热度相结合,这种结合既能利用启发式规则的不需要训练语料库就能生成符合语法规则的压缩句的优点,又通过保留了尽量多的热词。本文算法提出的启发式规则简单易于实现,不需

要太多的语言学技巧。因为中文的“原句-压缩句”对齐语料库较难获得，本文不使用对齐语料库的特点降低了实验的难度。词语热度的引入使生成的压缩句在损失较少的压缩率和语法性的前提下尽量保留较多的热词，加大了压缩句的热度，可以为用户保留更多最新最热的信息。

接下来的工作中会做一些算法方面的改进，将会改进本文的启发式规则，使用一些更通用精准的规则减少语法方面的错误，再考虑优化启发式规则与热度的结合方式，使之在不降低语法性和信息量的前提下尽量多地保留热词。改进后的系统将会用来自动生成文摘和标题。

参考文献：

[1] Jing H.Sentence reduction for automatic text summarization[C]//Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, WA, USA, 2000: 310-315.

[2] Zajic D, Dorr B, Lin J.Single-document and multi-document summarization techniques for email threads using sentence compression[J].Information Processing and Management, 2008, 44(4): 1600-1610.

[3] 沈剑虹.RSS：信息整合传播的未来[J].河北大学学报：哲学社会科学版，2006，31(2)：133-135.

[4] Corston-Oliver S.Text compaction for display on very small screens[C]//Proceedings of the NAACL Workshop on Automatic Summarization (WAS 2001), Pittsburgh, PA, USA, 2001: 89-98.

[5] Knight K, Marcu D.Summarization beyond sentence extraction: a probabilistic approach to sentence compression[J]. Artificial Intelligence, 2005, 139: 91-107.

[6] Nguyen L, Shimazu A, Horiguchi S, et al.Probabilistic sentence reduction using support vector machines[C]//Proceedings of 20th COLING, Switzerland, 2004: 743-749.

[7] McDonald R.Discriminative sentence compression with soft syntactic constraints[C]//Proceedings of 11th EACL, Trento, 2006: 297-304.

[8] Hori C, Furui S.Speech summarization: an approach through word extraction and a method for evaluation[J].IEICE Transactions on Information and Systems, 2004, E87-D (1): 15-25.

[9] Turner J, Charniak E.Supervised and unsupervised learning for sentence compression[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005: 290-297.

[10] Clarke J, Lapata M.Global inference for sentence compression an integer linear programming approach[J]. Journal of Artificial Intelligence Research, 2008, 31: 399-429.

[11] Xu Wei, Grishman R.A parse-and-trim approach with information significance for Chinese sentence compression[C]//Proceedings of the 2009 Workshop on Language Generation and Summarisation, 2009: 48-55.

[12] 赵青.基于概率统计和句法分析的中文语句压缩系统的研究与实现[D].北京：北京邮电大学，2012.

[13] Dorr B, Zajic D, Schwartz R.Hedge trimmer: a parse-and-trim approach to headline generation[C]//Proceedings of HLT-NAACL, Text Summarization Workshop, Edmonton, 2003, 5: 1-8.

[14] Li W, Xu W, Wu M, et al.Extractive summarization using inter-and intra-event relevance[C]//Proceedings of COLING ACL 2006, Sydney, 2006: 369-376.

[15] 邢文训，谢金星.现代优化计算方法[M].北京：清华大学出版社，1999.

（上接39页）

[6] 王凯.时间依赖网络中国邮路问题的分支限界算法[D].大连：大连理工大学，2008.

[7] Gillet B E, Miller L R.A heuristic algorithm for the vehicle dispatch problem[J].Operations Research, 1974, 21: 340-349.

[8] AoBrand A, Jos A A.A new tabu search algorithm for the vehicle routing problem with Backhauls[J].European Journal of Operational Research, 2006, 173(2): 540-555.

[9] 吴洁明.物流配送车辆路径优化问题的仿真研究[J].计算机仿真，2011，28(7)：357-360.

[10] 王铁君，邬月春.基于混沌粒子群算法的物流配送路径优化[J].计算机工程与应用，2011，47(29)：218-221.

[11] 罗勇，陈治亚.基于改进遗传算法的物流配送路径优化[J].系统工程，2012，30(8)：118-122.

[12] Low C, Chong Y, Salleh K H, et al.Path optimization using genetic algorithm evolution[J].IEEE, 2010, 13/14(12): 252-255.

[13] 温金保，蔡延光.基于自适应小生境遗传算法的物流配送路径优化研究[J].广东工业大学学报，2011，28(3)：20-23.

[14] 周艳聪，孙晓晨，余伟翔.基于改进遗传算法的物流配送路径优化研究[J].计算机工程与科学，2012，34(10)：118-122.

[15] 王华东，李巍.粒子群算法的物流配送路径优化研究[J].计算机仿真，2012，29(5)：243-246.

[16] 王宇平.进化计算的理论和方法[M].北京：科学出版社，2011.