基于组块的中文自动文摘系统研究

Research on the Chinese Automatic Abstracting System Based on Chunk

索红光 曹淑英 （中国石油大学(华东) 计算机与通信工程学院 山东东营 257061）

摘要：传统的基于统计的自动文摘方法以词语作为文本信息的基本单位，没有考虑到词语在不同语言环境下的具体语义，导致文摘精度不高。为了克服传统方法的缺点，提出了一种基于文本组块的自动文摘方法。系统利用中科院的 ICTCLAS 软件对文档进行分词和词性标注，并根据一系列的规则，将相关的词语构造成组块。由句子中出现的组块作为衡量句子重要性的标准选出文摘句。文中给出了自动文摘的评价方法和实验结果，跟传统的基于词语的文摘相比较，实验结果表明基于文本组块的自动文摘系统生成的文摘句精度更高，更能全面反映原文的主要内容。

关键词：自然语言处理 文本组块 自动文摘 统计方法 向量空间模型

1 引言

自动文摘研究如何利用计算机自动地从自然语言文本中提取摘要(摘要应包含原文的核心内容或用户感兴趣的内容)，并以语义连贯的段落乃至篇章的形式输出。

自 20 世纪 50 年代末，由 Luhn [1] 首先提出了自动文摘的概念并设计了一个自动文摘系统以来，自动文摘的方法和技术一直在不断的发展变化中，特别是进入 90 年代后，大量电子信息的涌现促使人们更加充分的认识到自动文摘的价值，自动文摘的研究进入了前所未有的繁荣期。

国内从 80 年代末开始研究自动文摘实验系统，至今为止，研究人员们已经进行了大量的研究并取得了一定的成果。但目前的技术水平尚不成熟，问题主要是在中文本身的语言特点以及自然语言理解方面的困难。

在当前，自动文摘系统大体上可分为两类：一是基于理解的文摘，利用句法和语义知识或者一阶谓词逻辑，对文本的内容进行分析；在理解的基础上，自动生成句子，形成摘要。但对于中文摘要来说，很难取得理想的文摘，这是因为汉语真正负载信息的是词而不是字，需要对文本进行分词处理。汉语的词汇极其丰富，同一个概念可以用很多不同的词汇表达，同一个词也可能根据语言环境有不同的意义。在句子构造上，与西文相比汉语的语法尚未形成规范化，而且人们习惯于非规范化的语法，对中文文章的理解还有很多问题。因此，基于理解的摘要方法在非受限领域短期内很难取得理想的结果。二是基于统计的文摘。基于统计的文本摘要方法主要根据线索词典、词频、词或句子的启发性函数进行模式匹配，摘取文本中重要句子形成摘要。它不依赖于具体领域，适应面广，响应速度快，但是往往单纯地依赖于词频，缺乏语义的支持。

结合基于理解的摘要方法与基于统计的摘要方法思想，进行了基于文本组块的自动文摘方法研究。下面将在第 2 部分中介绍基于 ICTCLAS 分词的组块处理和向量空间模型建立，在第 3 部分介绍组块重要度计算，第 4 部分介绍基于组块的 VSM 自动文摘生成，在 5、6 部分分别介绍系统评估以及最后的结论。

2 基于 ICTCLAS 词性标注的组块构建和向量空间模型建立

利用中科院提供的 ICTCLAS 软件对文本进行分词和词性标注，在此基础上进一步构造文本组块。

2.1 组块简介

组块是一种结构，是符合一定句法功能的基本短语。每个组块都有一个核心词，并围绕核心词展开，以核心词作为组块的开始或结束[2]。

在对自然语言理解的研究过程中，针对困难复杂

的句法分析，许多研究人员尝试着把一个完整的句法分析问题分解为几个易于处理的子问题，以逐步降低完整句法分析的难度，提高分析效率，这种方法称作组块分析。相对于完全句法分析，组块分析不再着眼于分析整个句子的语法和主题，而是仅仅把句子解析成较小的具有独立意义的单元，并不揭示这些单元之间的句法关系。

在语言中，具有实在意义的只是几种基本类型的组块，这几种组块是构成语言的基本元素。根据中文特点，组块类型可以划分为以下几种：

（1）基本副词短语（BADVP）：一般是以副词为核心词充当副词功能的短语。

（2）基本形容词短语（BADJP）：指核心词为形容词或充当形容词功能的短语。

（3）基本数量短语（BMP）：表示数量的短语。

（4）基本处所短语（BNP）：表示的是基本短语中的表示地点、地域或者表示这些概念的名词结构。

（5）基本名词短语（BNP）：是一类紧密结合的名词结构，由修饰词 + 名词核心词构成。

（6）基本动词短语（BVP）：包括动趋搭配、动补搭配、形式动词加实意动词等。

（7）其他（OP）：包括叹词、语气助词等。

2.2 基于简单规则的组块划分和概念向量空间模型建立

2.2.1 预处理

利用中国科学院计算所提供的汉语词法分析系统 ICTCLAS 对文本进行分词处理，按照词语在文中出现的顺序进行标注记录。

2.2.2 组块划分

汉语的组块划分有很多种方法，有基于最大熵模型方法[3]、基于统计的方法[4]、基于神经元网络的方法[5]、基于 SVM 的方法[6]、基于 Cotraining 机器学习的方法[7]等。

由于本文的实验系统对组块的划分质量要求不是很严格，因此，采用基于规则的方法对词语进行组块划分和处理。通过制订一系列简单规则，根据规则对有词性标注的词语进行初步处理组合，形成需要的组块。根据本文第 2 部分提到的几种组块基本类型，对文本进行组块划分。具体根据以下几个规则完成。

（1）若有“和”“与”“跟”等连词，判断该词两边的词是否是同一种词性，如果词性相同，则合并成一个短语，短语词性与连词两边的词词性相同。例如，连词两边同时为名词，则将该连词与两个名词合并成一个短语，为基本名词短语（BNP）。

（2）若有多个名词相邻，将所有相邻的名词合并成为一个基本名词短语（BNP）。

（3）若有多个形容词相邻，则将其合并形成一个基本形容词短语（BADJP）。

（4）将所有形容词跟其后面相邻的名词合并，作为基本名词短语（BNP）。

（5）若有“在”“于”等介词，后面是时间或处所以及与这些概念相关的名词，则将介词与后面的名词搭配形成基本处所或时间短语（BNP）。

（6）若数词之后的词为量词，则合并成一个基本数量词短语（BMP）。

（7）若有多个动词相邻，则合成为一个基本动词短语（BVP）。

（8）若有多个副词相邻，则合成为一个基本副词短语（BADVP）。

（9）若副词或形容词之后的词为动词，则跟动词合并成基本动词短语（BVP）。

经过组块划分后，文本的表示单位已由词语变成组块。采用计算每个未登录组块在文章中出现的次数是否大于预先制定的阈值，如果是则标注为另外一种新的组块；否则，删除那些在文章中出现次数小于阈值的未登录组块。

2.2.3 构造组块向量空间模型

组块划分完成后，在构造空间向量模型时，使用组块向量空间模型，而不是简单的以词语为单位的向量空间模型；即 VSM 由以前的 $S_i(W_1,F_{1i};\ W_2,F_{2i};\ W_3,F_{3i};\ \cdots\cdots W_n,F_{ni})$ 变为现在的 $S_i(C_1,F_{1i};\ C_2,F_{2i};\ C_3,F_{3i};\ \cdots\cdots C_n,F_{ki})$

（其中 $k\leqslant n$），C_i 为相互独立的一个个组块，而 F_{ij} 为互不相同的组块在 S_i 中出现的频度。

3 组块重要度计算

基于 2.2 组块向量空间模型建立的算法，完成对所有词语进行组块构建后，得到一个组块集合 $\{C_1, C_2, C_3, \cdots\cdots C_n\}$，统计组块在句子中出现的频度 F_{ij}，定义文本的组块向量空间模型。

[定义 1] 对已得到一个组块集合{ C_1, C_2, C_3, ……C_n}的文本，定义其组块向量空间模型为：

$$C=(F_{ij})_{n\times m}\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1m} \\ F_{21} & F_{22} & \cdots & F_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ F_{n1} & F_{n2} & \cdots & F_{nm} \end{bmatrix}$$

其中 n 为文本中的组块个数，m 为文本中句子个数，第 i 行第 j 列元素 F_{ij}表示组块 C_i在句子 S_j中出现的频度。

为了计算每个句子的重要度，系统还计算出各组块 C_i的重要度 $W(C_i)$。其中 F_{ij}为句子 S_j中组块 C_i的出现频率。则 $W(C_i)$ 的计算公式如下：

$$W(C_i)=\lambda\times\log\sum_{j=1}^{m}F_{ij} \quad (1)$$

其中 F_{ij}表示组块 C_i在句子 S_j中出现的频度，λ 是当 C_i 为标题组块的加权系数，在本系统中是以 1.2 作为参数计算的。

4 基于组块 VSM 的自动文摘生成

4.1 句子重要度计算

句子重要度计算是基于进行组块构建处理后的文本建立起句子的向量空间模型 S_j(C_1,F_{1j}; C_2,F_{2j}; C_3,F_{3j};……C_n,F_{kj})进行句子重要度计算。经过对大量文本试验结果进行分析后发现，句子权重主要与句子所包含的组块、句子本身所处段落的位置等因素密切相关，设计句子权重的计算函数为：

$$W(S_j)=\lambda\frac{\sum_{i=1}^{n}F_{ij}\times W(C_i)}{MM'} \quad (2)$$

其中 $W(C_i)$ 为 C_i的重要度，F_{ij}表示组块 C_i在句子 S_j中出现的频度，M 为句子 S_j包含的所有组块数，M' 表示句子 S_j 中所包含的分句数。λ 为当前句子是段落的句首或者是结尾时的加权值，本系统设为 1.3。

4.2 减小文摘冗余度

通过上面的步骤可以得到一个粗略的文摘，但往往出现一个问题就是文摘句的冗余度比较大。这是因为为了起强调作用，某些句子可能会在文章的不同位置反复出现，而这些句子都很重要，容易同时选入文摘，造成内容重复。为了消除冗余，还要进行句子相似度计算，删除掉语义重复的冗余句子。

[定义 2] ∀ 句子 S1、S2，用 SameWC(S1,S2) 表示 S1 和 S2 中相同概念的个数。则句子 S1,S2 的语义相似度为：

$$Sim(S1,S2)=2\times\frac{SameWC(S1,S2)}{Len(S1)+Len(S2)} \quad (3)$$

其中，Len(S1) 和 Len(S1) 分别是句子 S1、S2 中具有的组块个数。

对每个抽取出来的句子之间的相似度进行计算，如果相似度大于 0.7，则认为 S1、S2 描述的是同一个主题，选择重要度大的一个句子作为文摘句，删除重要度较小的句子。降低冗余度后，形成一个较好的文摘，然后进一步加工形成最后文摘。

5 系统评估

实验中，分别用从人民网上选取的 50 篇新闻类和 50 篇叙事类文章和自 CNKI 论文库中选取的 50 篇科技论文作为测试语料进行实验测试，分别用基于词语的自动文摘方法和基于组块的自动文摘方法进行文摘并对结果进行比较。

如何评价摘要的质量，目前仍然没有一个客观准确的衡量标准。即使是人工摘要也很少能达到唯一性，对于同一篇文献可以有若干篇可以被接受的摘要，不同的文摘员或者同一个文摘员在不同的时间内均可能写出不同的摘要。在实验中，我们参考文献[8]的评价方法采用主观评价的方法请三位专家分别对文摘结果进行人工判定，综合考虑文摘的可读性、安全性、概括性、准确性等各项因素对摘要进行打分，取其平均分进行比较。采用 5 分制来评价摘要质量，标准如下：5 分，如果摘要全面反映文本内容，语句通畅；4 分，能够反映文本内容，句子语义完整；3 分，基本概括了文本内容，但句子不够理想；2 分，对文本内容概括不够全面，只有部分内同；1 分，偏离文本主题，句子语义不完整。

其中，“长度”一栏表示抽取的文摘占待处理文本长度的百分比。从表中可以看出，对各类题材的文章，特别是新闻类和科技类的文章，基于组块的自动文摘方法比基于词语统计的自动文摘方法有明显的优势。

表 1 两种方法专家评分结果比较

体裁	方法	长度					
		5%	10%	15%	20%	25%	30%
新闻类	基于词语	2.10	2.53	3.03	3.60	4.06	4.20
	基于组块	2.47	2.80	3.50	4.03	4.33	4.66
科技类	基于词语	1.93	2.43	2.90	3.46	3.66	3.93
	基于组块	2.23	2.63	3.00	3.56	3.86	4.23
叙事类	基于词语	2.20	2.46	2.80	3.23	3.43	3.60
	基于组块	2.43	2.76	3.00	3.46	3.83	4.03

经分析,是因为具有一定语义独立性的组块比简单的词语更能代表文章意义,同时也避免了一些对主题不是很重要的高频词的干扰。例如,在一篇谈台湾问题的文章中,"台湾"与"中国"都是出现频率极高的词,其出现频率远远高出其它词,经过组块处理后"台湾问题""台湾海峡""中国领土""中国政府"等出现较多的组块大大减小了"台湾"与"中国"两词的频率,避免了因这两个极高频词的出现而造成"台湾处于紧紧环绕中国的岛链上,它是中国直面太平洋的最前沿"这种不是主题句的句子权重偏大的问题。

6 结束语

本文给出的基于组块的中文自动文摘方法,利用有独立语法意义的组块作为基本单位对文本进行处理,比单纯使用词语作为基本单位对文本进行处理更有意义,选出的句子准确率更高。

目前实验中发现的问题在于对于组块的划分研究还不够充分,有些组块由于条件的限制,没有很好的划分出来,这是下一步要进行的工作。

参考文献

1 Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information [J], IBM Journal, 1957 10(1):309 –317.

2 李素建、刘群,汉语组块的定义和获取[A],收录于孙茂松、陈群秀主编,语言计算与基于内容的文本处理,全国计算语言学联合学术会议(SWCL2003)论文集[C],清华大学出版社,2003,110 –115.

3 李素建、刘群、杨志峰,基于最大熵模型的组块分析[J],计算机学报,2003,31(17):3 –5.

4 刘芳、赵铁军、于浩等,基于统计的汉语组块分析[J],计算机学报,2003,14(6):28 –32.

5 王荣波、池哲儒,基于神经元网络的汉语组块自动划分[J],计算机工程,2004,30(20):13 –135.

6 李珩、朱靖波、姚天顺,基于 SVM 的中文组块分析[J],中文信息学报,2003,18(2):1 –7.

7 刘世岳、李珩、张俐等,Cotraining 机器学习方法在中文组块识别中的应用[J],中文信息学报,2004,19(3):73 –79.

8 林鸿飞、高仁璟,基于潜在语义索引的文本摘要方法[J],大连理工大学学报,2001,41(6):744 –748.