基于词汇链分析的英文自动文摘

Automatic text summarization using lexical chains

管鹏玲 刘贵全 （中国科学技术大学计算机科学技术系 安徽合肥 230027）

摘要：提出并实现了一种改进的基于词汇链分析的自动摘要系统，该系统结合语法分析的特点计算候选词的上下文窗口，提高了采用贪心策略构建词汇链的质量；并针对传统的词汇链分析方法生成的摘要长度的相对固定性给予了改善。实验表明，该系统与采用原始贪心策略构建词汇链的自动摘要系统相比较，生成的文摘质量有明显提高。

关键词：自动摘要 词汇链 句法分析 依存关系 上下文窗口

1 引言

摘要是保留原文信息、把原文压缩为更精炼的1文摘的过程[1]。

目前主要的摘要技术有三类：基于浅层分析的方法、基于话语结构的方法、基于实体分析的方法[2]。基于浅层分析的方法是建立在文本表层的形式特征基础上的，缺乏对文本内容的深层次分析，难以保证生成文摘的逻辑连贯性，文摘质量的提高受到了限制。基于话语结构的方法易受文档结构影响，对于结构不规范的文档，摘要效果就很差。基于实体分析的方法先分析出文本内部的概念性表示，然后提取出文档中各实体并建立起实体间的关系，据此来确定各实体对表述文档内容的作用。这种方法中采用最多的是词汇链分析方法。

2 词汇链分析方法

正如 Halliday 和 Hasan[4] 所阐述的那样，一篇文档的句子及其词汇具有统一性，往往是描述同一些事物的，这些句子通过回指、联接词及词之间的语义关系形成某种凝聚力。其中，由单词之间的语义关系而产生的这种凝聚力称为词汇凝聚力[3-5]。国外很多研究者通过分析词汇凝聚力来构建词汇链[4]，把词汇链作为文档的中间表示，最后形成摘要。通常，这种方法可描述为以下几个步骤[6]：

（1）选择一系列候选词，候选词通常是名词或名词短语；

（2）对每一个候选词，依据关系准则[6]在链集合中找到一条合适的链，把词插入到链中，如果没有合适的链，则为该候选词创建一条新链；

（3）对构建完的词汇链，分析计算出强词汇链[7]，即最能表示文档核心内容的词汇链；

（4）对每条强词汇链，根据一定规则，抽取具有代表性的句子，重组润色后形成摘要。

多数研究者采用这种方法很大的原因在于词汇凝聚力是一种易于识别的关系，可使用 WordNet[10-11] 等词典作为构建词汇链的主要知识库。此外，词汇链为文本结构和主题的确定提供了重要线索，方便形成文档的概念化表示，从而把握文档的主要意思。

构建词汇链可采用动态或贪心的策略[7][9]，采用动态策略构建词汇链的方法对于每个候选词，考虑它在 WordNet2 中的所有意思，分别加入到相应的链中，所有候选词处理完之后，计算出互不冲突的若干条强词汇链。这种方法需要指数级的内存空间，对于大规模的文档集，这个问题显得尤为突出。

贪心策略的主要思想是从问题的某一个初始解出发，通过一系列的贪心选择——当前状态下的最优选择，逐步逼近给定的目标，以尽可能快的求得更好的解。当达到算法中的某一步不能再继续前进时，算法停止。

采用贪心策略构建词汇链的方法认为一个词出现在文中某个位置，它与附近某些单词之间存在着某种联系，对每个候选词，根据上下文，分析得出其在文中

唯一确定的意思，然后加入到某条唯一的词汇链中。这样就一步步求解得到每个候选词的意思，得到局部最优解，逐渐逼近全局最优解，即所有候选词的意思。

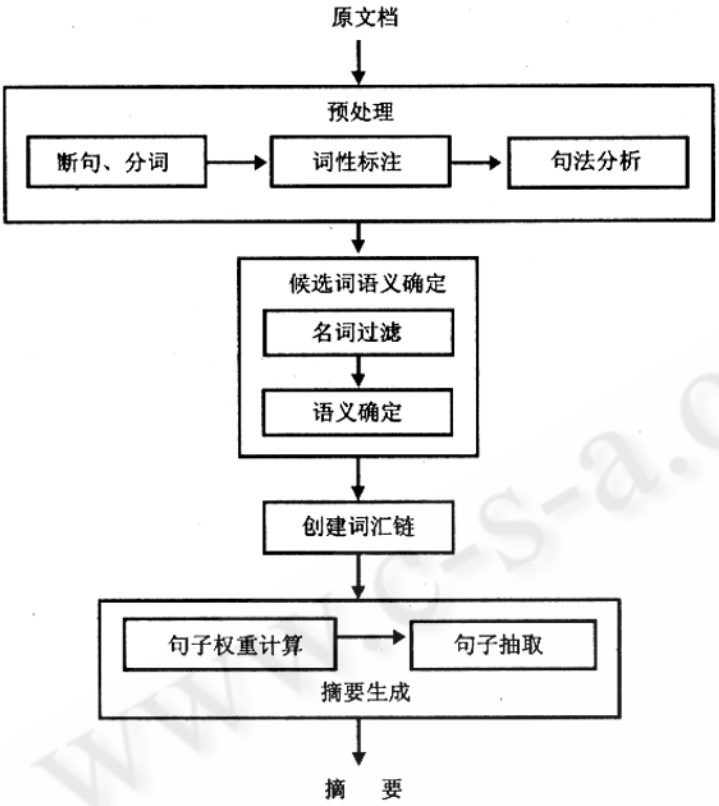

图 1　系统流程图

有研究者提出用候选词前后几个单词在 WordNet 中解释的相互交集，来确定它们的意思。实际上，一个单词只和上下文中某几个单词意思之间存在某种联系，如果考虑某些不相干的词，则可能会干扰单词在文中意思的确定。选择哪些合适单词来参与候选词意思的确定，是本文将要讨论的一个问题。

此外，原始的基于词汇链分析的自动摘要方法在摘要生成阶段抽取句子时，对每条强词汇链只抽取一条句子，不考虑文档的长度和实际摘要所需比率。如何确定合适的摘要长度，选择有意义的句子作为摘要也是本文将要讨论的问题。

3　算法描述

在总结现有方法的基础上，本文提出了一种改进的基于贪心策略构建词汇链的方法，即对每个候选词，根据上下文，分析得出其在文中唯一确定的意思，然后加入到某条唯一的词汇链中。而不是对所有的词进行分析处理、分别加入到多个词汇链之后，再判断词的意思。与传统的采用贪心策略的词汇链方法不同，我们在分析候选词的意思时，不是直接拿上下文中几个单词在 WordNet 中的解释互相参照来确定候选词的意思，而是在对句子进行句法分析之后，确定和候选词有关联的某些实词，根据这些有关联的实词来共同确定候选词的意思。此外，在摘要生成阶段，不是简单地从每个强词汇链中抽取一条代表性的句子，而是对每个句子的权重依据一定的准则进行计算后，根据需求抽取可变数量的句子数。

系统的流程图如图 1 所示。

3.1　预处理

首先利用 Lingpipe3 提供的接口对文档进行断句、分词、词性标注，过滤停用词。由于名词包含了文档的主要意思，我们根据词性分析的结果，将名词作为候选词加入到词汇链中。在将名词（候选词）加入到词汇链之前，需要确定该候选词的意思，以便判断其与哪条词汇链中的词有某种关联，这样才可以加到包含这些关联词的词汇链中。为了确定候选词的意思，我们首先对做好词性标注的句子进行句法结构分析，根据句法分析结果得到的句法结构来判断和候选词具有关联的词。然后根据这些有关联的词及候选词在 WordNet 中的解释来确定候选词的意思。下面将详细阐述。

常见的句法结构形式有句法树、依存关系树（依存语法[11]、范畴语法）、有向图（链语法）、特征结构（HPSG、LFG）等等。我们采用了 Dekang Lin 的英文依存分析工具 Minipar4，该工具采用的是依存语法的语法体系（dependency grammar）。依存语法中的依存关系满足如下五条公理：

（1）一个句子只有一个成分是独立的；

（2）其他成分直接依存于某个成分；

（3）任何一个成分都不能依存两个或以上的成分；

（4）如果 A 成分直接依存于 B，而 C 成分处于 A，B 之间，那么 C 或者直接依存于 A，或者直接依存于 B，或者直接依存于 A 和 B 之间的某一个部分。

（5）中心成分左右两边的其它成分相互不发生关

系。

采用这种句法分析将句子由一个线性序列转化为一棵结构化的依存分析树，通过依存弧反映句子中词汇之间的依存关系，树中不含有非终结符，一个词数为 n 的句子对应的依存树只有 n 个节点，n－1 条边，表示形式很简洁，易于操作，。如句子“He likes walking on the bank of lake.”的句法分析结果如图 2 所示。

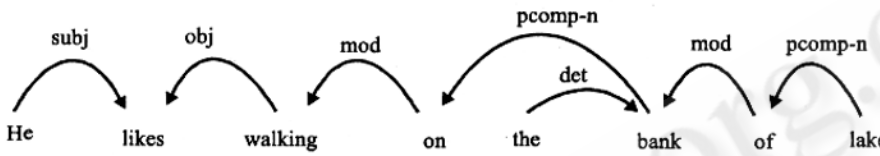

图 2　依存分析树

在图 2 中，如果两个词之间有弧相连，表示两者之间存在依存关系，弧发起的词（依存词）依存于弧指向的词（核心词）。弧上的标记表示该依存关系，如：词 He 依存于词 likes，依存关系为词 He 是词 likes 的主语（subj）。

3.2 确定候选词语义

根据句法分析的结果，对每一个候选词，我们定义它的上下文相关词集，称之为上下文窗口。以下是确定上下文窗口的算法：

（1）如果词 word 的某个子节点词 $word_i$ 是一个实词，则选取节点词 $word_i$，返回；否则，递归直至在 word 的子树中找到一个实词或者整个子树遍历结束。

（2）往词 word 的父节点方向寻找，如果父节点词 $word_i$ 是一个实词，则选取节点词 $word_i$，返回；否则，继续往根节点方向寻找，直至找到一个实词为止，然后返回；

依存语法认为发生依存关系的一对词中，如果词 A 修饰词 B，则 B 为主词，A 为从属词，A 是 B 的附属成分，在依存关系树中体现为 A 是 B 的子结点，故我们在算法中首先考虑候选词的子节点，修饰成分一般为形容词、名词居多；再向父节点方向寻找支配该候选词的词，一般为动词、动名词居多。

通过此算法确定出候选词 word 的上下文窗口后，再将上下文窗口中每个单词 wi 与候选词 word 在 WordNet 中各自不同解释进行比较，计算 w_i 和 word 的解释中包含的相同词的最大个数，称之为释义关联度 PR（Paraphrase Relation）。设 w_i 在 WordNet 中有多个解释，P_{wi} 为其中一个解释，，同样，word 在 WordNet 中也有多个解释，P_{word} 为其中一个解释，将 P_{wi} 和 P_{word} 分别表示为词集 $P_{wi}=\{t_1,t_2,\cdots,t_n\}$，$P_{word}=\{t_1,t_2,\cdots,t_m\}$，其中 $t_i(1\leqslant i\leqslant n)$ 为 P_{wi} 中包含的词，$t_j(1\leqslant j\leqslant m)$ 为 P_{word} 中包含的词，n 和 m 分别为 P_{wi} 和 P_{word} 中词的个数，则 P_{wi} 和 P_{word} 这两个解释的释义关联度 PR 为：

$$PR(P_{w_i},P_{word})=\sum_{i=1}^{n}\sum_{j=1}^{m}I(t_i,t_j)$$

其中，$I(t_i,t_j)=\begin{cases}1\ldots t_i=t_j\\0\ldots t_i\neq t_j\end{cases}$。

计算 w_i 和 word 在 WordNet 中所有满足词性（根据预处理阶段词性分析的结果）的对应解释的关联度后，选取 w_i 和 word 的释义关联度最大的对应解释分别作为两单词的意思。举例如下：根据上下文窗口计算方法我们可以得到图 2 中各候选词的上下文窗口，如表 1 所示。

表 1　图 2 中结点的上下文窗口

结点	上下文窗口
walking	like, bank
bank	walk, lake
lake	bank

对于名词 bank，其在 WordNet 中有多种解释，其中最主要的两种分别是：

bank1：a financial institution that accepts deposits and channels the money into lending activities；

bank2：sloping land（especially the slope beside a body of water.

而名词 lake 的解释有：

lake1：a body of（usually fresh）water surrounded by land

lake2：a purplish red pigment prepared from lac or cochineal

得到各种解释间的释义关联度 PR 为：

PR（bank1，lake1）＝1；

PR（bank1，lake2）＝1；

PR（bank2，lake1）＝5；

PR(bank2,lake2)=1;

故选取释义关联度最大的一组解释(bank2,lake1)作为词 bank 和 lake 的解释。若存在多个解释对的释义关联度相同,则选取包含在 WordNet 中排序靠前的释义的解释对。以上这种分析词义的方法,避免了过多词参与候选词意思的确定,减少了干扰信息,在长句子中这种优势体现地更为明显。

3.3 构建词汇链

对每个候选词,根据前阶段获得的意思,依据关系准则[6],和链中的词进行关系计算,判断它们是否具有特别强关系(重复)、强关系(同义反义、上下义、部分整体关系等)、中介关系[6],这些关系可直接通过词在 WordNet 中的关系来判断。若在某条链中找到和候选词具有这些关系中的任一种关系的词,则把候选词加入到该链中。若在任何链中都找不到合适的词与候选词具备这任一种关系,则为该候选词新创建一条链,并把它添加进去。

3.4 计算词汇链的权重

根据如下公式[7]计算词汇链的权重

$$Score(Chain)=\sum_{m=1}^{k}w_m * H \quad (1)$$

$$H=1-\frac{K}{\sum_{m=1}^{k}w_m} \quad (2)$$

$$Score(Chain)>Average(scores)+2*StandardDeviation(Scores) \quad (3)$$

其中,w_m 为链中第 m 个词在文中出现的频数,K 为链中成员数,H 是一个均一性指数。满足公式 3 的为强词汇链,其中,Average(scores)为所有词汇链的平均值,StandardDeviation(Scores)为标准偏差。

3.5 摘要生成

不同于传统的词汇链摘要方法,我们首先根据多种信息对文档中的句子计算权重,接着依据权重进行排序后,根据文档长度和摘要率计算得出的摘要长度,抽取一定数量的高权重句子。

3.5.1 句子权重计算

句子权值的计算公式为:

句子权重 $=\lambda_1$ * 单词的权重 W $+\lambda_2$ * 句子中单词所属链的权重 S $+\lambda_3$ * 句子中概念的跨度 CS $+\lambda_4$ * 位置信息参数 L $+\lambda_5$ * 其他信息参数 I　(4)

这里采用单词在文中的词频作为权重 W,是对预先过滤了停用词后的高频词的照顾,认为高频词作为文档信息的一种载体,在一定程度上反映了文档的内容。

由于强词汇链是文档中较重要概念的一种表示,在计算词汇链阶段已被赋予较大的权值,因此将句中单词所属链的权重 S(采用 3.4 节的计算方法)作为计算句子权重的一个因子。

本文考虑了句中单词所属链的范围,即该句中所有候选词所属链的集合,又称句子的概念跨度 CS。若句子中所有候选词所属链的个数(句子的概念跨度值)大于一个阀值 F,我们就认为该句的信息量跨度比较大,需要赋予较大的权值。

在其他信息参数 I 中,我们考虑每个强词汇链的代表性词所属的第一个句子,额外地给这些句子的权重一个奖励 I,经过多次实验,本系统最终取值 I=1.5,(实验方法同下文 λ_i 的值调整)这就保留了传统词汇链方法抽取句子的优势。此外,我们还考虑了句子的位置信息,对段首段尾的句子,给予一些特殊照顾,加重它们的权值 L。

公式中的 $\lambda_i(1\leqslant i\leqslant 5)$ 分别是各权重的加权系数。我们选取了 10 篇不同的文档,首先将每篇文档中所有句子按重要程度人工排好序,并给 λ_i 设定一个初始值,分别对每篇文档中所有句子按公式(4)计算出的权重值进行排序,判断其排序结果是否跟人工给出的排序结果一致,反复调整 λ_i 的值,从中选出一个相对最优的结果(即公式(4)计算出的句子权重排序结果与人工排序结果一致性最大)。在以后的工作中,将考虑如何自动确定最优的加权系数。

3.5.2 抽取句子

各句的权值计算出来后,将各句依其权值排序。根据文档的长度和摘要的比率来确定所需的句子数,依次将权值最大的句子选入文摘,直到文摘达到特定长度。最后依据句子在原文中的顺序排序,得到最终摘要。

这种方法避免了传统该类方法摘要句子数目的固定性,能根据实际需求来抽取可变长度的摘要,提高了摘要的灵活性。

4 实验结果和评估

本文采用的是一种内部评价方法,对系统进行定

量评测。首先从网络上收集100篇不同风格的新闻文摘作为测试语料；每篇文章由两位专家各自独立地做出手工文摘，以此作为“理想”文摘。其中，每位专家给每篇文章做两次摘要，每次文摘地长度分别为原文长度的20%和30%，文章长度以句子数来计算。将本文提出的自动摘要系统跟理想摘要进行比较，通过计算平均精确率(Precision)和召回率(Recall)来评价系统生成文摘的质量。精确率和召回率按如下公式计算：

$$Precision = \frac{|S_l \cap S_m|}{|S_m|} \quad (5)$$

$$Recall = \frac{|S_m \cap S_c|}{|S_c|} \quad (6)$$

其中S_m是系统生成文摘的句子集，S_l是理想摘要的并集，S_c是理想摘要的交集，算子“|||”取集合的势。

由于基于原始贪心策略构建词汇链方法形成的文摘是固定长度的，而本文提出的摘要系统(a)可以根据用户需求进行可变长度的摘要，故我们在原始方法的摘要生成阶段也采用本文中提出的方法，将两者进行比较，得到的系统性能评价结果如表2所示。

表 2　系统性能评价

	摘要比率	系统(a)	系统(b)
20%	精确率	0.742	0.653
	召回率	0.761	0.695
30%	精确率	0.75	0.674
	召回率	0.784	0.714

由表2可见，与基于原始贪心策略构建词汇链方法的摘要系统比较，本文提出的摘要系统在精确率和召回率上均有明显提高，且在一定程度上弥补了传统词汇链方法的缺陷。

5　结语

本文提出并实现了一种基于贪心策略创建词汇链方法的英文摘要系统，在原始的方法上添加了句法分析，以便更准确地判断候选词的意思，正确的创建词汇链。此外，还结合了多种方法对基于词汇链分析方法的文摘系统增添了灵活性，同时，也提高了生成文摘的质量。

参考文献

1 Karen Sparck Jones. What might be in summary [R]? Proceedings of Information Retrieval 93: Von der Modelherung zur Anwendung, Universitatsverlag Knstanz, 1993, 9－26.

2 Mani I., Maybury M. Advances in automatic text summarization[M]. Cambridge: MIT Press,1999. Ⅰ－Ⅶ.

3 Morris J., Hirst G. Lexical Cohesion computed by thesaural relations as an indicator of the structure of text[J]. Computational Linguistics, 1991. 17(1).

4 Michael Hasan and Ruqaiya Halliday. Cohesion in English[M]. London: Longman,1976.

5 Hasan R. Coherence and Cohesive Harmony[C]. In J. Flood(ed), Understanding Reading Comprehension. Delware: International Reading Association,1984.

6 Graeme Hirst and David St－Onge. Lexical chains as representation of context for the detection and correction of malapropisms[M]. In Christiane Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambridge, MA: The MIT Press, 1997.

7 Barzilay R. Lexical Chains for Summarization[D]. Master's thesis, Ben－Gurion University, Beer－Sheva, Israel. 1997.

8 Brunn M., Chali Y., and Pinchak C. Text summarization using lexical chains[C], In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001, New Orleans, Louisiana, 2001.

9 Pedersen T, Banerjee S., and Patwardhan S. Maximizing semantic relatedness to perform word sense disambiguation[R]. University of Minnesota. Supercomputing Institute,Research report UMSI 2005/25.

10 George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An online lexical database [J]. International Journal of Lexicography (special issue), 1990,3(4):235－312.

11 Tesniere L. Element de Syntaxe Structurals[M]. Paris: Klincksieck, 1959.