云南大学学报（自然科学版），2003，25（3）：217~220　　CN 53-1045/N　ISSN 0258-7971
Journal of Yunnan University

基于实时联想的医学诊断报告书语言生成器*

施心陵[1]，　王　逍[1]，张榆锋[1]，汪源源[2]
（1.云南大学 电子工程系，云南 昆明　650091；2.复旦大学 电子工程系，上海　200433）

摘要：在某一特定的领域中，组成自然语言句子的成分之间存在着很强的相互联系，根据这种相互联系可以由句子的中心语联想产生出句子的其他组成成分.据此，提出了一种基于实时联想的自然语言句子的生成方法，给出了一个医学诊断报告书语言生成器.

关键词：联想频率；基于实时联想；句子生成；有限词汇联想储存模型
中图分类号：TP 18　　文献标识码：A　　文章编号：0258-7971(2003)03-0217-04

自然语言句子的构成，在某个特定的领域中，一般说来都是在中心语的基础上加上各种修饰成和补充说明成分而构成的，组成句子的各部分词语之间存在着很强的优先组合关系[1].据此，鉴于医学诊断报告书常用语言一般限制在一个有限的范围内，本文提出了一种实用的基于实时联想的医学报告书语言的生成方法，医学诊断报告书语言句子各部分组成成分之间具有密切的相互联系[2]，分析词语间的相互联系，利用联想的方法，本文实现了一个基于实时联想的医学诊断报告书语言生成器.经过一定的训练，在数据库中储存一定数量的医学词汇之后，可以不用输入或很少输入文字来帮助医生生成诊断报告书.

1　医学诊断报告书有限词汇的联想储存结构

要计算机来生成自然语言句子，首先要建立一种结构，以表达出需要交流的信息.本文以 Findler 和 Hendrix[3,4] 提出的联想网络为基础，利用特定领域有关知识之间的相互联系，构造一个动态联想的有限词汇存储模型.

使用分块联想网络来表示动态联想网络有限词汇储存模型.将相互联系的有关词汇组织成森林状的层次分类结构，结点代表存储的词汇.对于同类型的词汇，用树结构来组织，构成动态联想网络存储模型中的一个超结点.树中的孩子结点是双亲结点派生（联想）出来的子知识项的集合.联想储存模型如图 1 所示，图中有 A，B，C 3 个超结点.以超结点为根，可以不断地向下派生出和上一层父结点联系的下层结点.各超结点之间通过共有的结点相互联系，通过这些共有的结点，在超结点之间建立起了相互联系，组成了一片相互联系的森林.

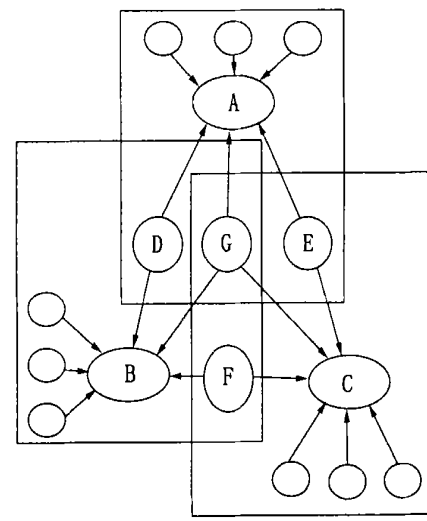

图 1　有限词汇的联想存储模型

Fig.1 Limite vocabulary associative memory network

* 收稿日期：2003-03-27
基金项目：云南省自然科学基金资助项目（2002C0002Z）；云南省省院省校合作项目.
作者简介：施心陵（1956-　），男，教授，主要从事信息分析、信息处理、人工智能方面的研究.

连接结点的弧定义为由双亲结点到孩子结点的联想概率. 本文使用信息论中互信息[5]来体现从双亲结点到孩子结点之间的联想关系的大小程度.

根据医学诊断报告书句子特点和句子的各部分成分在句子中所起的作用和所处的位置和解决生成句子时的联想的方便,及为了医学词汇储存的简洁并尽量减少储存词汇的冗余量,对句子的 8 种成分构造如下的 3 个超结点来进行储存,各个超结点之间通过中心词为共有结点相互联系. 使用关系型数据库来实现联想储存,结构如下:

中心词超结点储存结构

Basic _part	Basic _method	Spell	Headword
诊断部位	诊断方法	汉语拼音	中心词

句子其它成分超结点储存结构

Headword	Other _part	Attribute	Frequency	Condition	Incept
中心词	句子其它成分	属性	联想频率	使用概率	起始概率

句子修饰成分超结点储存结构

Headword	Modifier _part	Frequency	Condition	Incept
中心词	句子修饰成分	联想频率	使用概率	起始概率

对 3 个超结点储存结构作如下的分析:

中心语超结点储存句子的主语,是基于联想的句子生成的起点,句子的修饰成分和句子的其它成分都是由中心语作为起点来进行联想得到的. 为了快速、方便地查找,在中心词表中定义了 2 个字段诊断部位和诊断方法来对该储存结构进行限制.

句子的其它成分超结点储存句子的谓语、宾语和补语. 对于特定的应用环境来说,一定中心语后面的句子的其它成分应该大体上是一样的,所以对句子的其它成分储存结构没有用任何附加的条件来进行限制. 在句子其它成分超结点中,定义了一个属性的字段,该字段是用来表明所储存的其它成分是谓语动词还是句子的后接成分.

句子的修饰成分超结点储存除了句子的主语、谓语、宾语和补语之外的剩余的句子的所有成分. 对于特定的应用环境来说,一定中心词的修饰用语应该大体上是一样的,所以对句子的修饰成分储存结构没有用任何附加的条件来进行限制.

其它成分和修饰成分中的使用概率和起始概率是用余计算词语之间的联想概率的,下一节将进行详细的说明.

在上面的 3 个储存结构中,通过中心语共有结点,就把 3 个超结点的树结构组成了一片相互联系的具有层次森林结构.

2 基于实时联想的医学诊断报告书句子生成方法

词 A, B 之间的联想概率,应用信息论和概率统计中互信息的概念有

$$I(A,B)=\log_2\left[\frac{P(A,B)}{P(A)P(B)}\right],$$

而

$$P(A,B)=P(B/A)P(A).$$

代入词 A, B 的联想概率的公式中,有

$$I(A,B)=\log_2(P(B/A)/P(B)),$$

条件概率 $P(B/A)$ 和概率 $P(B)$ 的计算,做如下的规定:

对于每一个词语,储存的时候进行一个该词语的储存次数的统计,称为该词语的起始次数. 在使用词语的过程中,每使用到 1 次,则它的使用次数就加 1,得到的词语总的次数为使用次数. 对于每一个词语,要计算它和中心语之间的互信息,则需要得到它本身出现的概率和由中心语确定的条件概率 2 个数值. 词语本身的概率值,使用统计所得到的起始次数来计算的,词语的起始次数是固定的,则词语本身的概率值也是固定的;对于词语的条件概率值,由词语的使用次数来计算,由于词语的使用次数是变化的,则词语的条件概率值也是变化的. 反映到词语的互信息公式中,则词语由中心

语所想起的互信息(联想频率)的数值也是变化的, 由词语被中心语联想得到的使用次数来确定, 是实时的.

得到了联想概率后, 就可以来进行句子的联想生成了. 采用有穷自动机[7](简称 FA)来表示句子的生成过程. 对于有穷自动机中的 5 元组 $M = (Q, \Sigma, \delta, q_0, F)$, 定义如下:

非空有穷状态集合 Q: 所有用于生成句子的词语的集合;

非空有穷输入字母表 Σ: 所要生成的句子的词语的集合;

状态转移函数 δ: δ_1: 联想; δ_2: 词语的选择; δ_3: 句子生成;

初始状态 q_0: 用于联想的整个句子中心语;

终结状态集合 F: 所生成的句子.

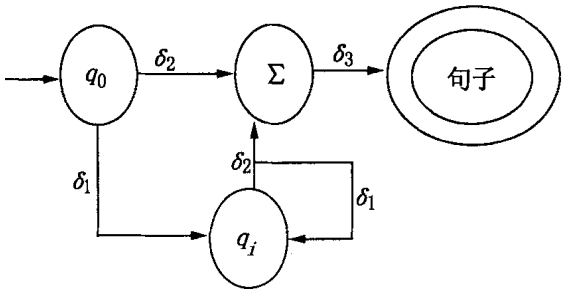

图 **2**　句子联想状态转移图

Fig. 2 A state move graph for a sentence using their association

如图 2, 选择句子的中心语 q_0 到句子成分集合 Σ, 再由 q_0 联想得到 q_i, 从 q_i 中选择所需要的部分到 Σ 中, 进行句子的生成. 如果需要, 以 q_i 中选择的句子的组成成分为中心语, 根据互信息的大小, 来联想句子的下一步的组成成分, 再从联想所得到的句子的组成成分中选择所需要的部分来进行句子的生成. 还可以再进行下一步的联想, 这样不断地进行, 直到完成句子的联想生成. 对于状态转移函数 δ_1, δ_2, δ_3 的条件, 是由使用者来进行控制的.

3　医学报告书语言生成器

根据上面的基于实时联想的自然语言句子的生成方法, 使用 Visual C++, 实现了一个医学报告书的语言生成器.

在程序中, 分别实现了一个句子成分分析器和句子生成器. 句子成分分析器用于对诊断报告书进行分析, 把词汇按照分类添加到相应的超结点储存结构中. 句子生成器则是通过联想, 从储存结构中获得句子的词语, 组合词语来生成句子.

句子成分分析器是用来对句子生成器进行学习训练的. 本文以医学影像学诊断报告书书写手册[2]一书中的诊断报告为例, 对 40 篇医学报告书进行了学习训练, 向医学诊断报告书有限词汇的储存结构中填充了 1 739 个医学词汇. 以后的任何时候, 都可以使用分析器来进行学习、训练.

句子生成器的程序以句子的中心语为起点, 使用词语之间的联想概率, 来联想出句子的各个部分的组成成分, 选择所需要的句子的组成成分后, 来生成句子. 在联想得到句子的各部分组成成分的时候, 生成器同时进行词语之间联想概率的调整, 反映词语之间实时的联系. 通过一个例子来说明它的工作过程.

拟生成的诊断报告为: "两肺清晰. 心脏、横隔及片上显示肋骨均未见异常.", 生成过程为:

(1) 获取整个句子的中心语 q_0. 先为句子的中心语选择限制条件. 选择诊断部位: 胸部. 选择诊断方法: 平片. 经过查找, 得到在该诊断部位和诊断方法的限制下 q_0: 胸部.

(2) 由 q_0 联想得到 q_i. 以 q_0 为中心语, 经过联想, q_i 的动词成分有: 未见, 显示, 建议, 可见. 后接成分: 心脏, 肺, 胸片, 纵隔, CT 检查, 横隔, 胸腔, 气管, 肺野, 心. 由于生成句子所需要的句子组成成分已经得到, q_i 后接成分仅联想出了和 q_0 之间互信息最大的前 10 项. 修饰成分: 及, 至. 如图 3 所示.

从 q_i 的后接成分中选择肺为句子的组成成分 Σ, 然后在以肺为中心语联想, 得到后接成分: 清晰, 癌肿, 上叶, 结核, 层面, 肿块. 动词成分: 未见, 伴, 可见, 受压. 修饰成分: 两, 右, 上, 之, 此外, 左. 从肺的后接成分中选择清晰为句子组成成分 Σ, 从肺的修饰成分中选择两为句子的组成成分 Σ, 现在 Σ 中的词语为: 肺, 清晰, 两. 则完成了第一句话生成: 两肺清晰.

(3) 依次类推, 完成诊断报告书中剩余句子的联想生成. 从 q_i 的后接成分中得到心脏, 横隔, 修饰成分中得到及, 动词成分中得到显示. 以显示为中心语联想, 从显示的后接成分中得到肋骨, 以肋

骨为中心语联想, 从肋骨的动词成分中得到未见, 修饰成分中得到片上. 最后以未见为中心语联想, 从未见的后接成分中得到异常, 修饰成分中得到均, 则最终完成了句子: 心脏、横隔及片上显示肋骨均未见异常. 至此, 以整个句子的中心语头颅为起点 q_0, 使用各句子成分逐步的联想, 完成了诊断报告 F: "两肺清晰. 心脏、横隔及片上显示肋骨均未见异常." 的联想生成. 如图 4 所示.

图 3 q_0 联想成分

Fig. 3 q_0 association components

图 4 诊断报告完成

Fig. 4 A medical imaging diagnostic report is finished

4 小 结

本文提出了一种基于实时联想的自然语言句子的生成方法, 并使用这种方法实现了一个医学诊断报告书句子生成器. 生成器生成的句子是可控的, 可以产生任意长度和任意复杂度的句子, 可由使用者来控制. 因此生成的自然语言句子的结构接近于自然. 在储存数据库中储存了 40 篇诊断报告书的情况下, 使用该生成器来生成所储存的报告书, 可以完全由计算机联想得到报告书的所有组成成分. 使用该生成器生成其它类别的报告书, 则有一部分词语不能联想得到, 需要手工输入. 这个问题可以通过学习、训练生成器来解决.

基于联想的自然语言句子的生成, 适用于句子各部分成分之间相互联系紧密的情况, 而且句子的语法结构应尽可能的简单. 如果句子的结构很复杂, 句子各部分成分之间的联系不是很紧密, 则在基于联想的自然语言句子的生成中, 如何实现有效的联想以及如何设计联想的知识存储结构等等方面, 还有很多的问题有待进一步地深入研究.

参考文献:

[1] 沈天真, 陈星荣. 医学影像学诊断报告书书写手册[M]. 上海: 上海医科大学出版社, 1996.

[2] 吴蔚天, 罗建林. 汉语计算机语言学——汉语形式语法和形式分析[M]. 北京: 电子工业出版社, 1994.

[3] FINDLER N V(ed). Associative networks: representation and use of knowledge by computers[M]. New York: Academic Press, 1979.

[4] HENDRIX G G. Encoding knowledge in partitioned networks[A]. Findler N·V·(ed)·Associative networks: representation and use of knowledge by computers[C]. New York: Academic Press, 1979, 51—92.

[5] 常 迥. 信息理论基础[M]. 北京: 清华大学出版社, 1993.

[6] 王克宏, 汤志忠, 胡 蓬, 等. 知识工程与知识处理系统[M]. 北京: 清华大学出版社, 1996.

[7] 王 逍, 张俊华, 施心陵. 一种基于联想的自然语言句子的生成方法[J]. 昆明理工大学学报, 2001, 126: 209—202.

[8] 王 逍, 陈建华, 李 甦, 等. 基于动态联想网络的有限词汇储存结构[J]. 计算机应用, 2002, 10: 36—37.

(下转第 224 页)

of turing machines[J]. J Stat Phys, 1982, 29: 515.
[32] BENIOFF P. Quantum mechanical hamiltonian models of turing machines that dissipates no energy[J]. Phys Rev Let, 1982, 48: 1581.
[33] BOQOMOLNY E B. Semiclassical quantization of multidimensional systems[J]. Nonlinearity, 1992, 5: 805—866.
[34] 郝柏林. 生物信息学手册[M]. 上海: 上海科学技术出版社, 2003.
[35] 彭守礼. 探索生命的遗传语言[A]. 李喜先. 21 世纪的 100 个科学难题[C]. 长春: 吉林人民出版社, 1998. 435—439.

Symbolic dynamics, star product and others

PENG Shou-li
(Center for Nonlinear Complex Systems, Department of Physics, Kunming 650091, China)

Abstract: The chaos break down the classical Laplace determinism[1]. Hence it changes the viewpoint in which people look closely at the world, the symbolic dynamics provides a mathematical representation of this viewpoint.

Key word: symbolic dynamics; star product; chaos; laplace determinism; universality

* *

(上接第 220 页)

A language generator of medical imaging diagnostic report based on real-time association

SHI Xin-ling[1], WANG Xiao[1], ZHANG Yu-feng[1], WANG Yuan-yuan[2]
(1. Department of Electronical Engineering, Yunnan University, Kunming 650091, China;
2. Department of Electronical Engineering, Fudan University, Shanghai 200433, China)

Abstract: There are strong mutual relations among the components of natural language sentences in a given field. According to these relations, we can generate other components of a sentence using their association by the sentence's key component. Hereby, A method for generating natural language sentences based on real-time association is presented and a language generator of medical imaging diagnostic report is given.

Key words: associative probability; based on real-time association; generating sentence; limite vocabulary associative memory network