一个简单人机对话系统的实现方法

胡凤国

（中国社会科学院研究生院语言系，北京 100102）

E-MAIL: bushiwoshishui@yahoo.com.cn

摘要： 本文主要谈的是我们在进行一次人机对话实验过程中的实际做法和遇到的困难，以及采取的解决办法。在内容上详细介绍了这个实验模型的实现策略，在某些特定环节上提出了自己的观点。实验中对 HL 模型进行了开放测试，共实验了 100 个句子，正确率 92%，结果基本上是令人满意的。

关键词： 人机对话；句式变换；模式匹配；句法分析

引言

人机对话包括语音对话和书面对话两个方面，本文主要说的是书面对话。国外开展人机对话研究较早，早期使用关键词和模式匹配的方法，后来引入句法分析、语义分析和知识推理，建成了不少对话系统，其中比较著名的有伍兹（W. Woods）的 LUNAR 系统、维诺格拉德（T. Winograd）的 SHRDLU 系统等。国内在这方面的研究起步比较晚，但近年来随着 IR 技术和 IE 技术的出现，人机对话研究在国内得到了越来越多的重视。

为了深入探讨基于模式匹配的隐模板系统（伪理解系统）的缺陷，以便提出改进意见，为进一步开发智能理解系统做好理论上和实践上的准备，我们以作家作品等文学常识为题材进行了一次人机对话实验，在故意不参考先贤著作的情况下独立探索，建立了一个实验模型（HL 模型）。本文详细介绍了 HL 模型的实现策略，在某些特定环节上提出了自己的观点。实验中对 HL 模型进行了开放测试，共实验了 100 个句子，正确率 92%，结果基本上是令人满意的。

1 HL 模型的设计原则和实现方法

总的设计原则是：首先按照类型把问题分成五种句式，然后针对不同的句式使用不同的处理方法。对于每一种句式，都要先根据切词词典进行切词，然后进行必要的句式变换，在此基础上再进行句式合并，最后得到有限的几种简单句式。通过上面的处理，将问题逐步简化，最后很容易地得到结果。

1.1 句式分类

根据作家作品人物之类文学常识的习惯问话方式，在充分考虑日常问话习惯的基础上，我们发现，绝大多数问题，都可以归结为以下五种类型：

1） 一般疑问句

一般而言，这类问句的特征是最后一个汉字为“吗”而且句中有“是”、“有”等关键字，或者句末没有“吗”但句中有“是不是”、“有没有”等标志。

2） 特殊疑问句

这一类句子末尾没有“吗”，在实际判断中，凡不具备其他四种句式特征的都归入此类。

3） 选择疑问句

这一类句子末尾也没有“吗”，它跟特殊疑问句的区别是：句中一般有“……是……还是……”这样的关键词。

4） 合取主语疑问句

这类句子不但末尾有“吗”，而且句中还有“……和/跟/与……都……”等关键字。可分解成两个一般疑问句。例如：

✧ 《茶馆》和《日出》都是老舍写的吗?

5） 选择主语疑问句

这类句子，句末没有“吗”，但是句中有“……和/跟/与……谁/哪……”等标志。可分解成两个一般疑问句。例如：

✧ 《屈原》和《顽主》里面，哪一部是王朔的作品?

1.2 句子切词

1.2.1 切词词典

切词词典是由作品表、人物表这两个数据库以及系统的保留字表组建而成的。作品表和人物表的内容主要是一些常识性的数据，系统保留字表主要存放的是对句式判别起着决定作用的字和词语。切词词典中的每条记录分为两部分：词本身以及该词的属性。如果词本身是保留字，那么其属性已经在保留字表里面给出，如果词本身是数据库中的数据，那么其属性就是它在数据库中对应字段的属性。部分词及其属性举例如下：

金庸 作者　　林黛玉 人物　　人物 关键词　　有 动词　　谁 疑问词　　和 连词

1.2.2 切词算法

HL 模型的切词算法是正向最大匹配。这里没有采取复杂的分词算法主要是考虑到文学常识类的问句一般比较简单，极少产生歧义。

1.2.3 未登录词处理

在未登录词的处理上，HL 模型同一般的分词系统有所不同。第一、一般的分词系统，对于未登录词，通常是切成单字；而在 HL 模型中，则是把未登录词忽略掉。第二、一般的分词系统，把作品名中的未登录词与一般未登录词同等看待，仍是切分成单字； HL 模型在切分时，强制把左右书名号中间的部分切到一起，并给它赋予“作品”的属性，不管它是不是存在于切词词典中。这样处理未登录词，有利于留住句子的主要部分，便于以后的句式变换和句式合并。例如：

A) 【句子】: 王熙凤是《红楼》中的人物吗?
　 【切分】: 是 红楼 的 人物 吗 ?

B) 【句子】: 项少龙在其中出现的那部作品的作者是谁?
　 【切分】: 项少龙 的 作品 的 作者 是 谁 ?

1.3 句式变换

同一个问题，不同的问法是很多的，据统计，询问作品的作者，所有可能问法竟然有上百种之多。为了方便计算机对问句的“理解”，我们设置了句式变换，把含义相同的问句映射成与之等价的基本句式。

HL 模型的句式变换原则有两个：一是尽量往基本句式的方向靠拢；二是使具有领属关系的“AB”往

“A 的 B”这个方向靠拢。目前使用的句式变换规则有十余个，兹举一例：

【规则】：　　作者 + “作品”　　→　　作者 + “的” + “作品”

【变换】：　　金庸 作品 有 哪些 ？　　→　　金庸 的 作品 有 哪些 ？

此外，还有两条特殊的规则用来变换双主语比较格式的问句。所谓“双主语比较格式”的问句，表面上看不属于一般疑问句，但经过变换，能化为一般疑问句，举一例如下：

【句子】：　　林黛玉和王熙凤是同一个人写的吗?

【切词】：　　林黛玉 和 王熙凤 是 同一 人 写 的 吗 ？

【变换】：　　林黛玉 的 作者 是 王熙凤 的 作者 吗 ？

1.4 句式合并

在句式变换的基础上，把相邻的“A 的 B”结构合并成一项，以此来简化句式。合并的时候，我们根据已知的项“A”，到数据库中寻找它的属性为“B”对应项。假设合并得到的结果为“D”，那么，“D”的属性应该置为“B”。只要能够操作，这种合并可以连续进行，例如：“A + 的 + B + 的 + C”当中，“A + 的 + B”先合并，得到“D”，然后，“D + 的 + C”还可以再执行合并操作。举一例说明：

【合并前】：项少龙 的 作品 的 作者　　（注：“项少龙”这一项的属性为“人物”）

【合并中】：寻秦记 的 作者　　（注：“寻秦记”这一项的属性为“作品”）

【合并后】：黄易　　（注：“黄易”这一项的属性为“作者”）

1.5 寻找答案

一个句子经过切词、变换、合并之后，已经是一个非常简化的基本句式。针对各种不同的句式类型，必须采取不同的处理办法。

1.5.1 一般疑问句

一般情况下，一般疑问句能化成一个基本句式：“A + 是/是不是/有/有没有 + B + ……”。这可分为三种情况来进行处理：

第一种：“A”是单项，“B”是单项：直接比较字符串“A”和 “B”的值，相等则返回“是的”，否则返回“不是的”。例如：

【句子】：诸葛亮是男的吗?

【切词、变换、合并】：男 是 男 吗 ？

【答案】：是的.

第二种：“A”是单项，“B”是集合：处理方法是把“B”拆分成一个个单项，分别跟“A”进行单项对单项的比较，只要有一个比较结果为“是的”，那么整个句子结果为“是的”，否则整个句子就回答“不是的”。例如：

【句子】：《四世同堂》是鲁迅的作品吗 ？

【切词、变换、合并】：四世同堂 是 阿Q正传 祝福 药 吗 ？

【答案】：不是的.

第三种：“A”是集合，“B”是单项：处理方法是把“A”拆分成一个个单项，分别跟“B”进行单项对单项的比较，只要有一个比较结果为“是的”，那么整个句子结果为“是的”，否则整个句子就回答“不是的”。例如：

【句子】：　鲁迅的作品中有《四世同堂》吗?

【切词、变换、合并】：阿Q正传 祝福 药 有 四世同堂 吗 ？

【答案】：没有.

1.5.2 特殊疑问句

1） 基本句式的求解

符合下列三种句式之一者，为基本句式，均以A作为问题答案而直接返回：

A) A + ?

B) A + “是”/“有” + “谁”/“哪” + ?

C) “谁”/“哪” + “是”/“有” + A + ?

2） 扩展句式的求解

但凡不是基本句式的都属于扩展句式。为扩展句式寻找答案，要先找已知和未知。已知就是适合充当“A的B”结构中的“A”项的词语；未知就是适合充当“A的B”结构中的“B”项的词语。已知，一般情况下，它的属性是“作者”、“作品”、“人物”中的一个，而且一个简化后的句子中只会出现一个，我们很容易就把符合这样条件的项找出来。寻找未知就比较麻烦，它有两种办法：

第一、 直接办法：句中有下面两种模式之一的，就可以直接断定未知应该是“作者”。

【模式1】 “谁” + 动词

【模式2】 “谁” + “的” + “作品”/“著作”/文体

第二、 间接办法：这是一个经验方法，选取距离疑问词最近的一个适合充当“B”的项。如果句中没有疑问词，就假定第一项为疑问词。

【句子】: 郭沫若话剧作品?

【切词、变换、合并】: 郭沫若 话剧 ? （据此确定“话剧”是未知项）

【答案】: 紫棣之花 屈原 .

1.5.3 选择疑问句

把这种句子分成两个分句，两个分句都是一般疑问句，调用一般疑问句的判定过程即可。看哪个分句的返回值为“是的”，则输出相应的部分。例如：

【句子】: 诸葛亮是女的还是男的?

【分解】: 诸葛亮是女的吗? （返回值“不是的”） 诸葛亮是男的吗? （返回值“是的”）

【答案】: 答案: 男的.

1.5.4 合取主语疑问句

把这种句子分成两个分句，两个分句都是一般疑问句，调用一般疑问句的判定过程即可。如果两个分句的返回值都为“是的”，则输出“对”，否则输出“不对”。例如：

【句子】: 《茶馆》和《日出》都是老舍写的吗?

【分解】: 《茶馆》是老舍写的吗? （答案: 是的.）

《日出》是老舍写的吗? （答案: 不是的.）

【答案】: 不对.

1.5.5 选择主语疑问句

把这种句子分成个分句，两个分句都是一般疑问句，首先调用一般疑问句的判定过程。然后根据两个分句返回值的不同作出相应的回答。这种句子的回答有四种方式，兹举一例说明：

【句子】: 林黛玉和诸葛亮哪一个是《三国演义》中的人物? 【答案】: 后者.

2 HL模型的性能评价

本系统用VC++6.0实现，采取批量问答方式进行操作。软件行业中主要用正确性和健壮性来表述一个软件的性能，具体到HL模型，正确性指的是它回答问题的正确率；而健壮性主要看它在错误数据输入的条件下的反应如何。

2.1 正确性

我们通过网上搜索和小范围的问卷，定了100个有代表性的问句进行了实验，系统回答无误的有92个。其他8个句子的错误类型主要有下面几种：

1、理解正确，回答时表达不是很确切（一个句子）：

【句子】：鲁迅写过《四世同堂》吗？　　【答案】：不是的。

2、对否定问句的回答模棱两可（一个句子）：

【句子】：《四世同堂》不是老舍写的吗？　　【答案】：是的。

3、问句中的主要成分是未登录词而造成理解错误（两个句子，举一例）：

【句子】：曹操和诸葛都是《三国演义》中的人物吗？【答案】：不能确定。

4、系统确实回答不出来（四个句子，举一例）：

【句子】：老舍写没写过《四世同堂》？　　【答案】：不知道。

2.2 健壮性

HL 系统主要考虑到输入数据的错别字/漏字问题、数据属性配套错误问题、句式省略问题以及数据不足问题，绝大部分情况下，系统能给出正确回答。

1、错别字/漏字问题：

【句子】：老蛇是《四世同堂》的作者吗？　　【答案】：不能确定。

2、数据属性配套错误问题

【句子】：四世同堂是《老舍》的作者吗？　　【答案】：对不起，不知道答案！

3、句式省略问题

【句子】：金庸作品？　　【答案】：笑傲江湖 鹿鼎记 射雕英雄传。

4、数据不足问题

【句子】：闰土是谁的作品里的人物？　　【答案】：不知道。

3 HL 模型的不足

HL 模型的不足之处主要有三点，这三点是我们以后对该模型改进时考虑的重点。

1） 程序跟数据没有分开

采用程序跟数据分开的好处是显而易见的，我们可以随时修改、添加规则而不用修改程序。目前，HL 系统还不能做到随时修改规则库。

2） 句式变换规则的数目太少

HL 模型还没有充分考虑到各种可能性，制定的规则过于简单，而且数量也偏少。如果增加几条规则的话，下面的问句就能很容易地回答出来。

- 老舍写没写过《四世同堂》？

3） 缺少语法分析和语义分析

HL 模型实际上是隐模板系统（伪理解系统），虽然问句中不存在让发问者填空的模板，但是模型内部仍然是按照模板的方式进行模式匹配，找到正确答案。乍看起来好像系统理解了问话，事实上这是个“伪理解系统”。这个伪理解系统只有切词，而没有语法分析和语义分析，虽然能对付绝大多数的问话，但是它仍然处理不了稍微复杂一点的问题。

4 改进设想

首先，增加模板数量。目前，HL 模型是基于大量的模板匹配进行工作的，如果增加模板的数量，那么就可能使系统对句式实行穷尽式的覆盖，从而提高系统的正确率。其次，建立学习机制，使系统能自动扩充数据库，自动总结句式变换规则。最后，从根本上改变 HL 模型的设计思路，引入语法语义语境语用分析和逻辑推理机制，朝智能理解系统的方向努力。

目前，我们正在参考先贤著作，改进 HL 模型，使之朝着智能理解的方向发展。闭门造车，思路难免受限，谨此为文，求教于方家，我们真诚希望能得到关于 HL 模型的任何批评和建议。

参考文献:

[1] 范继淹. 人机对话系列讲座. 语文战线, 1985, 2～5.

[2] 冯志伟. 自然语言的计算机处理. 上海外语教育出版社, 1996.

[3] 刘开瑛, 郭炳炎. 自然语言理解. 科学出版社, 1991.

致谢 中国社会科学院语言研究所傅爱平研究员和国家语言文字工作委员会冯志伟先生在 HL 模型的实现过程中给予了细心的指导，石锓博士对本文的完成提出了很多有益的建议，在此一并表示感谢。

作者简介: 胡凤国（1977－），男，山东曹县人，硕士生，主要研究领域为自然语言信息处理。

The Realization of a Simple QA-system

Hu Fengguo

(Graduate School of Chinese Academy of Social Science, Beijing 100102, China)

E-MAIL: bushiwoshishui@yahoo.com.cn

Abstract: QA-system is a challenging area of NLP and the research of which has made great progress at home and abroad. This article is meanly about what we had done with our experiment of a simple QA-system and the difficulty during the experiment, as well as the method to solve it. We introduced the realization strategy of this experimental model, and put forward our own opinion on some special step. The result of the test is acceptable with 92 right answers out of 100 questions.

Key words: QA-system; sentence pattern transformation; pattern match; syntactic analysis