

# 基于神经网络的问句组块分析

付 斌, 樊孝忠

(北京理工大学 计算机学院, 北京 100081)

**摘 要:** 问句分析是自动问答系统研究中的重点和难点。在中文问句的结构特点基础上, 结合机器学习及组块分析理论, 对问句进行组块分析, 实现了基于神经网络的问句组块识别算法, 并应用于银行领域自动问答系统中。测试结果表明, 对问句组块的识别能够达到比较满意的效果。

**关键词:** 自动问答; 组块分析; 语义块; 神经网络

**中图分类号:** TP181; TP183

**文献标识码:** A

**文章编号:** 1673-629X(2006)10-0094-03

## Question Chunk Analysis with Neural Networks

FU Bin, FAN Xiao-zhong

(School of Computer Science, Beijing Institute of Technology, Beijing 100081, China)

**Abstract:** Question analysis is not only important but also difficult task in question-answer (QA) system. Based on the structure feature of the question, machine learning and chunk parsing theory, an approach for question chunk parsing using neural networks is implemented. The approach is also used in question-answer system in bank field. Experiments prove that it is feasible to use the method for question chunk parsing.

**Key words:** question-answer; chunk parsing; semantic chunk; neural networks

### 0 引 言

一直以来, 由于自然语言的复杂性和不确定性, 完全句法分析是可望而不可及的。近年来一个发展趋势是由完全句法分析走向浅层句法分析, 即组块分析(chunk parsing)<sup>[1,2]</sup>。组块分析致力于识别句子中某些结构相对简单、但有重要意义的成分, 而不以完整的句法分析作为自己的目标。有些语言信息处理系统并不需要完整的句法分析, 而借助组块分析即可满足大部分的需求。自动问答系统中, 问句分析的目的是找出问句中对解答该问句起重要作用的信息, 这些信息可以表示成向量的形式来完整表达问句的语义。结合组块分析的特点和问句分析的目的, 将组块分析理论用于问句分析, 在问句实例研究的基础上, 将问句划分为不同语义块, 通过对这些语义块的自动识别来达到问句分析的目的。

用于汉语组块分析的机器学习方法有很多: 隐马尔科夫模型(HMM)<sup>[3]</sup>, 最大熵(maximum entropy)<sup>[4]</sup>, 支持向量机(SVM)<sup>[5]</sup>等。神经网络是一个学习分类器的有效方法, 该方法曾被成功地用于英语<sup>[6]</sup>和汉语<sup>[7]</sup>的短语边界识别。文中将神经网络的方法用于问句组块识别, 取得了不错的效果。

### 1 问句的组块分析

#### 1.1 基于 Chunk 的问句语义结构划分

Chunk(组块)是一种介于词汇和句子之间的、具有非递归特性的核心成分, 每个组块由句子中的单词或多词单元组成, 并且具有固定的语义。对问句进行组块分析, 首先要制定问句的组块类型。组块类型的确定和具体的研究有关, 有面向通用的划分, 也有面向具体领域的划分<sup>[8]</sup>, 无论哪种划分都应该遵循一定共同的原则: 组块不能破坏句子实际的句法结构; 组块的划分只依据局部的表层句法信息; 组块划分应能够使组块自动分析变得简单高效; 述宾和主谓不能出现在组块内; 为了降低后续的句法分析的难度, 一般使组块尽可能长一些。

文中的组块划分遵循了上述原则, 并结合了问句的结构特点和 QA 答案提取的需求。在定义问句组块前, 首先引入语义块的定义。

**定义 1:** 将句子中具有固定语义, 且其位置相对固定的部分称为语义块。

通过对 2000 句银行领域原始问句的分析, 总结出问句划分的 5 种语义块, 见表 1。

这 5 种语义块基本包含了问句的所有语义结构, 也构成了问句组块分析的标注集合。

一个经过组块分析后的问句如下所示:

例 1: 贵行提供了哪些低利率贷款

[En 贵行/n][Ev 提供/v 了/u][Qf 哪些/r 低/a 利率

收稿日期: 2005-12-11

**作者简介:** 付 斌(1980-), 男, 北京人, 硕士研究生, 研究方向为自然语言处理、自动问答系统; 樊孝忠, 教授, 博士生导师, 研究方向为自然语言处理技术, 网络教学技术。

/n 贷款/n]

表 1 问句语义块划分

语义块类型	名称
Av	属性值块
At	属性块
En	实体块
Ev	事件块
Qf	问点块

下面分别对主要的语义块做简要分析:

实体块(En): 实体块主要描述了问句中施事和受事主体, 主要是一些名词或名词短语。

问点块(Qf): 问点是问句询问信息的焦点, 问点块是对问点的完全描述。通常问点块由疑问词或由疑问词和相关的词结合而成。通过对标注好的问句实例的统计发现, 虽然不同类型问句的语义块的语义组成不同, 但是有一定的规则, 例如询问时间、数量值的组成规则如下:

Qf = 疑问词(什么|啥) + En 什么时间、啥时候等;

Qf = 疑问词(多少) + En 多少人、多少钱等;

属性块(At)和属性值块(Av): 属性块主要描述了实体的属性, 属性值块描述了实体属性的值。比如在例 1 中, ‘贷款’是一个实体(En), ‘利率’是‘贷款’的一个属性, ‘低’是‘利率’的值。由于‘低利率贷款’满足规则  $En = Av + At + En$ , 而  $Qf = \text{疑问词} + En$ , 因此‘哪些低利率贷款’被一起标注为问点块, 而没有分开标注。

事件块(Ev): 事件块主要描述了问句中实体的动作。例如例 1 中的‘提供了’是一个事件块, 它是实体块‘贵行’发出的动作。

## 1.2 问句组块实例分析

问句实例构成了问句实例库。问句实例库是后续机器学习方法的学习对象, 每个问句实例都经过了分词、词性标注和组块标注。其中分词和词性标注由计算机完成, 采用的是中科院张华平等的分词词性标注系统<sup>[9]</sup>, 并在原系统基础上加入了领域专业词库。组块标注主要由人工完成, 在组块标注的过程中初步形成了一个具有 85 条语义块构成的规则库。此规则库是对问句组块构成规律的初步总结, 对后续的人工标注和机器学习方法的验证具有指导意义。下面通过对一个原始问句的分析, 说明构造问句实例的全过程。

例 2: 什么人能够办理个人住房贷款

分词与词性标注: 什么/r 人/n 能够/v 办理/v 个人住房贷款/n

人工组块标注见图 1。

从图 1 可以看出几条组块规则, 比如  $Ev = v + v$ ,  $Qf = \text{疑问词} + n$ ,  $En = n$ 。

每一个原始句子经过上述两个步骤后成为一个问句实例, 最终形成 2000 句规模的问句实例库。

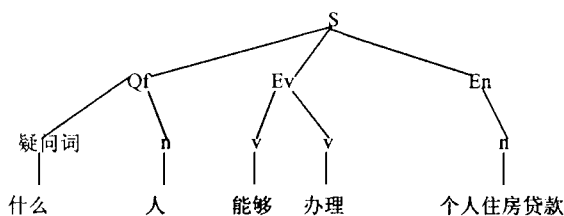


图 1 问句的组块标注过程示意图

## 2 数学模型的建立

通过对组块内词的相对位置进行编码, 可以将组块识别问题转化为一个标记问题, 由此就可以采用各种通用的分类模型。文中采用 BIO 编码方式, B 表示一个组块的开始, I 表示组块的内部, O 表示组块外的词, 如句子:

[Ev 具备/v][Qf 哪些/r 条件/n]的/u[En 单位/n]  
[Ev 才/d 可以/v 办理/v][En 信用卡/n]? /w

依据 BIO 可以编码为: 具备/B-Ev 哪些/B-Qf 条件/I-Ev 单位/B-En 才/B-Ev 可以/I-Ev 办理/I-Ev 信用卡/B-En ? /w

其中‘B-Ev’表示组块类型为 Ev 的第一个词, ‘I-Ev’表示组块类型 Ev 的内部词。B 和 O 同时也意味着前一个组块的结束。这样, 针对每一种组块类别有 B- 和 I- 两种标记, 如果有  $n$  种组块类别则对应  $2n + 1$  种组块的标记。组块识别问题就转化为一个  $2n + 1$  类分类器问题, 文中的组块定义为 5 种类型, 则组块识别问题可以转化为 11 类的分类器问题。

组块标注的对象是经过切分和词性标注的文本, 用  $S = (W, T)$  表示, 其中:  $W = (w_1 w_2 w_3 \cdots w_k)$  表示词序列,  $T = (t_1 t_2 t_3 \cdots t_k)$  表示对应的词性序列。另外, 用  $C = (c_1 c_2 c_3 \cdots c_k)$  表示文本  $S = (W, T)$  的组块标记序列。则对文本的组块标注问题转化为已知词序列  $W$  和词性序列  $T$ , 寻求概率最大的组块标记序列  $C$ , 使得:

$$C^* = \arg \max_C P(C | W, T) = \arg \max_C P(W, T | C)P(C) \quad (1)$$

假设词语和词类信息对组块标记的作用是独立的, 可得:

$$P(W, T | C) = P(W | C)P(T | C) \quad (2)$$

如果只考虑局部词语和词类信息对词语组块标记的影响, 可以简化得:

$$P(W, T | C) = \prod_{i=1}^n P(W_i | c_i)P(T_i | c_i) \quad (3)$$

其中  $W_i$  代表局部词语上下文信息,  $T_i$  代表局部词类上下文信息。

对于(1)式中的  $P(C)$ , 用二元模型可以简化为:

$$P(C) = \prod_{i=1}^n P(c_i | c_{i-1}) \quad (4)$$

综合以上 4 式, 可得:

$$C^* = \arg \max_C \prod_{i=1}^n P(W_i | c_i)P(T_i | c_i)P(c_i | c_{i-1}) \quad (5)$$

通过神经网络实现上述分类问题。神经网络具有比较强的非线性分类能力,同时又有比较强的抗干扰能力;神经网络的方法在上下文取词个数方面可以突破二元语法的限制,同时也不会引起空间的过度膨胀。

### 3 面向问句组块识别的神经网络结构设计

选取 BP 网络作为基本网型。BP 网是一种有指导学习的前向网络,能够学习复杂的非线性映射。

整个设计过程分为两部分:第一部分是网络的学习训练过程,即用训练语料对网络进行训练直至网络收敛,得到固定的连接权矩阵和阈值矩阵;第二部分是网络的回想过程,即运行已经训练完毕的网络对测试语料进行组块识别。我们在包含 2000 个问句的语料库中进行训练和测试,其中训练集含 1600 句,测试集含 400 句。

#### 3.1 输入层设计

本方法对输入的文本以词为单位进行处理。为了充分发挥上下文对组块标注的影响,输入层的设计不但要考虑当前词,也要考虑当前词的上下文环境。上下文环境包括 3 个方面:词语、词语的词性和词语的组块标记。因此将式(5)中的  $P(W_i | c_i)$ 、 $P(T_i | c_i)$  和  $P(c_i | c_{i-1})$  作为神经网络的输入特征。对于上述输入特征参数可以通过极大似然估计得到。对不同的输入特征分别设计输入节点。文献[7]中,作者实现短语边界识别选取了  $P(w_i | c_i)$  和  $P(T_i | c_i)$  两个特征。文中借鉴了其思想,并对特征的选取作了改进。首先,  $P(w_i | c_i)$  改为  $P(W_i | c_i)$ ,即在参数估计时加入了词汇上下文信息对组块标注的影响;其次,加入了特征  $P(c_i | c_{i-1})$ ,目的在于加入组块标记上下文的影响。下面分别讨论各个特征的输入节点的设计。

设  $L$  为当前词左边的词数,  $R$  为当前词右边的词数,  $|X|$  为组块标记数,  $|Y|$  为词类标记数。

$P(W_i | c_i)$ :考虑当前左边  $L$  个词和右边  $R$  个词的影响,每个节点可以用二维向量  $IW(i, k)$  表示。 $i$  对应词的位置,  $i = 0$  表示当前词;  $k$  表示第  $k$  种组块标记。共  $(L + 1 + R) * |X|$  个节点。 $IW(i, k)$  用统计量  $P(w_i | c_k)$  来代替。

$P(T_i | c_i)$ :考虑左边  $L$  个词和右边  $R$  个词的词性的影响,每个节点可以用三维向量  $IT(i, j, k)$  表示。 $i$  对应词的位置,  $i = 0$  表示当前词;  $j$  表示第  $j$  种词性;  $k$  表示第  $k$  种组块标记。共  $(L + 1 + R) * |X| * |Y|$  个节点。对于第  $i$  个词  $w_i$ ,可以分两种情况考虑:当  $j$  与  $w_i$  词性下标相同时,  $IT(i, j, k) = P(t_j | c_k)$ ;否则,  $IT(i, j, k) = 0$ 。

$P(c_i | c_{i-1})$ :每个节点可以表示为一个三维向量  $IC(i, j, k)$ 。 $i$  对应词的位置,  $i = 0$  表示当前词;  $j$  表示前一个词为第  $j$  种组块标记;  $k$  表示第  $k$  种组块标记。输入节点数为  $(L + 1 + R) * |X| * |X|$ 。对于当前词,  $i = 0$ ,因为还没有进行组块标注,设上一个词的组块标记下标为  $r$ ,若  $r = j$  则  $IC(0, j, k) = P(c_k | c_j)$ ,否则  $IC(0, j, k) = 0$ ;当  $i > 0$  时,所有的词语都没有标注,则  $IC(i, j, k) =$

$P(c_k | c_j)$ ;当  $i < 0$  时,所有的词语都经过标注,则当  $j$  和  $k$  同时与相邻的两个词语的组块标记下标相同时,  $IC(i, j, k) = 1$ ,否则  $IC(i, j, k) = 0$ 。

#### 3.2 输出层设计

BP 网络组块模型是一个 11 类的分类器,每一类代表一个 BIO 标记,具体见表 2。

表 2 BIO 组块标记集

标记类型	名称
B-Av	属性值块开始词
I-Av	属性值块中间词
B-At	属性块开始词
I-At	属性块中间词
B-En	实体块开始词
I-En	实体块中间词
B-Ev	事件块开始词
I-Ev	时间块中间词
B-Qf	问点块开始词
I-Qf	问点块中间词
0	组块外部词

因此输出层设计为 11 个神经元,每个神经元表示一个组块标记类型,对于一个输入(一个词),控制只有一个输出神经元被激活,即为该词的组块标记。

#### 3.3 隐含层设计

因为理论上没有确定隐含层节点数的一般方法,只有根据实验确定。

通过改变隐含层神经元的数量,分别进行训练和测试,最终选取隐含层个数为 50。以下为具体实验过程和结果。

### 4 实验与结果分析

在 1600 个句子的训练语料上进行了 3000 轮的训练,此时网络接近收敛。分别对 1600 个句子和 400 个句子进行封闭测试和开放测试,结果见表 3。

表 3 实验结果 ( $L = R = 2$ )

隐含层节点数	封闭测试正确率 (%)	开放测试正确率 (%)
20	74.46	70.83
30	80.32	75.29
40	86.88	82.21
50	90.15	85.22
60	88.78	83.02

分别对  $L = R = 1$ 、 $L = R = 2$  和  $L = R = 3$  进行实验,  $L = R = 2$  时实验效果最好。随着隐含层节点数的增加,正确率先呈现出升高的趋势,在 50 的时候达到最高,然后正确率有所回落。

### 5 结束语

基于神经网络的问句组块分析方法在自动问答系统

(下转第 100 页)

的容器控件,它们的透明度分别是 80% 和 100% (全透明),采用的基色是 `Qt::gray`。另外,在两个气泡容器控件中共画了 3 个 Label 控件,它们也被透明函数 `transparentize()` 处理了,其透明度分别是 80%、50% 和 100%。

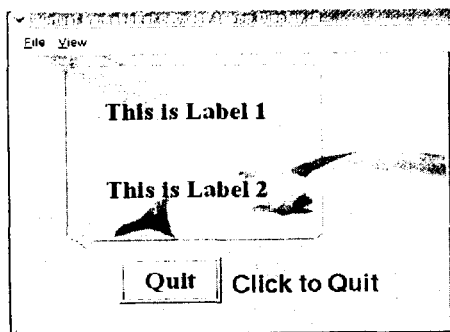


图 4 扩展控件的运行实例截图

### 3.3 控件合入 Qt/Embedded 类库

定制的控件完成后,可以把它合入 Qt/Embedded 类库,这样在以后需要使用时就可以当作 Qt/Embedded 基本控件直接使用了,从而提高工作效率。合入控件的方法比较简单,只需要在 `$QTEDIR/src/Makefile.in` 文件中增加控件的相应编译项就可以了。以本控件为例来进行描述,需要新增的内容如下:

(1) 控件的 moc 原文件,通过控件的头文件 (`tbubble.h`) 生成:

```
SRCMOC=moc-tbubble.cpp
```

(2) 对 moc 原文件编译生成的目标文件:

```
OBJMOC=moc-tbubble.o
```

(3) 控件的目标文件:

```
OBJECTS_ widgets=tbubble.o
```

(上接第 96 页)

中得到了应用,为生成问句向量提供信息来源,进而间接地为答案提取服务,效果良好,达到了预期目的。在后续的答案提取实验中答案匹配的准确率可达到 82.05%。然而在实验过程中也发现一些问题:在神经网络模型中,输入信息的提取十分重要,输入参数应该能够很好地反应问题的特征。本方法的一个不足是输入层的节点数目偏多,使得效率下降。对输入层的设计进行改进,使得输入层的节点减少,而又不会损失统计特征的特性,而且能够提高网络的效率和正确率。这是下一步要做的工作。

#### 参考文献:

- [1] Abney S. Parsing by chunks[A]. In: Berwick R, Abney S, Tenny C. Principle - Based Parsing[C]. Dordrecht: Kluwer Academic Publishers, 1991.
- [2] Abney S. Prosodic structure, performance structure and phrase structure[A]. In: Proc Speech and Natural Language Workshop[C]. San Mateo, CA: Morgan Kaufmann Publishers, 1992.

### 4 结束语

随着嵌入式技术的蓬勃发展,人们对嵌入式系统的图形用户界面也不断地提出新的需求,而现有的图形系统往往不能满足这一要求。笔者在对嵌入式图形系统 Qt/Embedded 的实现技术进行分析的基础上,成功地实现了个性化控件的定制,不仅满足了特定界面的需求,同时丰富了 Qt/Embedded 控件功能。

为了进一步定制结构复杂的控件以及实现动态的透明效果,今后还需做以下工作:继续研究 Qt/Embedded 控件绘制方法和事件驱动机制,在背景变化的情况下能实现动态的透明效果,从而定制出功能强大的扩展控件。

#### 参考文献:

- [1] Trolltech Inc. Qt Reference Documentation 3.0.6. 2002 [EB/OL]. <http://doc.trolltech.com/3.0/index.html>, 2002.
- [2] Trolltech Inc. Qt Reference Documentation 4.1.0. 2005 [EB/OL]. <http://doc.trolltech.com/4.1/index.html>, 2005.
- [3] Trolltech Inc. Qt/Embedded Whitepaper [EB/OL]. <http://www.trolltech.com/pdf/whitepapers/qt-embedded-whitepaper-a4.pdf>, 2002-03.
- [4] 徐广毅, 张晓林, 崔迎伟, 蒋文军. Qt/Embedded 在嵌入式 Linux 系统中的应用[J]. 单片机与嵌入式系统应用, 2004 (12): 14-18.
- [5] 吴伟清, 王磊, 吴朝晖. 基于 QTE 的嵌入式 Linux 中文环境解决方案[J]. 计算机工程, 2005, 31(2): 87-88.
- [6] 邹思铁. 嵌入式 Linux 设计与应用[M]. 北京: 清华大学出版社, 2002.

- [3] 李宏乔. 汉语组块分析技术及应用研究[D]. 北京: 北京理工大学计算机系, 2004.
- [4] 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, 26(12): 1722-1727.
- [5] 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, 2004, 18(2): 1-7.
- [6] Schmid H. Part-of-speech Tagging with Neural Networks [A]. In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94) [C]. Kyoto, Japan: [s.n.], 1994. 172-176.
- [7] 奚晨海, 孙茂松. 基于神经网络的汉语短语边界识别[J]. 中文信息学报, 2002, 16(2): 20-26.
- [8] 樊孝忠, 李宏乔, 李良富. 等. 银行领域汉语自动问答系统 BAQS 的研究与实现[J]. 北京理工大学学报, 2004. 24 (6): 528-532.
- [9] Zhang Huaping, Yu Hongkui, Xiong Deyi, et al. HHMM-Based Chinese lexical analyzer ICTCLAS [A]. 2nd SIGHAN workshop affiliated with 41st ACL [C]. Sapporo, Japan: [s.n.], 2003. 184-187.