基于模式匹配的自然语言识别

封春升　郝爱民
（北京航空航天大学计算机学院，北京 100083）
E-mail：fcsheng@263.net

摘　要　在自然语言识别过程中，为了提高识别的准确性，我们引入了模式匹配。不仅仅局限于传统的语法-语义分析，而是在语法分析的基础上，结合工程应用来定义最适合自然语言识别的语言模式，然后把模式存入到知识库当中。当需要对自然语言识别时，根据已有模式来匹配句子，从中检索出所需要的信息。文章完整地阐述了这种基于模式匹配的自然语言识别的全过程，并对模式的定义、分析及提取给出了详尽的剖析。最后以一个实验系统证明了此方法的可行性和准确性。

关键词　自然语言识别　自然语言处理　模式匹配　短消息处理

文章编号 1002-8331-(2006)19-0144-03　**文献标识码** A　**中图分类号** TP18

Automatic Recognition of Natural Language Based on Pattern Matching

Feng Chunsheng　Hao Aimin
(School of Computer Science & Engineering, Beihang University, Beijing 100083)

Abstract: In order to promote the accuracy in the recognition of natural language, we describe a method based on pattern matching. This method is not circumscribed in the analysis based on syntax-semantic, but considers both syntax and project application in order to define the proper pattern. We put all of the different patterns into knowledge database. When we need to recognize natural language, we match the natural language with the pattern, then pick up the useful information. This paper describes the integrated process of the recognition, and expatiates the pattern's definition, analysis and pick-up, at last, validates the feasibility and accuracy in the experimental system.

Keywords: natural language recognition, natural language processing, pattern matching, SMS processing

1　背景介绍

自然语言的处理是一个很庞大的工程，由于汉语语法的复杂性，到目前为止还没有一个非常有效的方法能够消除自然语言分析中的歧义问题。目前，自然语言的处理主要有以下几种方法：基于关键字匹配的方法；以句法-语义分析为主的方法；基于大规模语料库的自然语言处理。然而这几种方法由于自身的特点无法避免地都存在着弊端。基于关键字匹配的方法是一种近似匹配技术，主要的缺点是分析技术不精确，会导致很多的错误。以句法-语义分析为主的方法，分析起来很复杂。而基于大规模语料库的自然语言处理，则是一个庞大的工程，不适合于工程应用。随着自然语言的处理越来越趋向于实用化和工程化，我们必须提供一种高效准确的方法来识别自然语言。为此我们提出了一种基于模式匹配的自然语言处理方法，它能够处理任何一个特定领域的自然语言。比如下面提到的短信业务，往往只是针对某一个特定领域的。

短信服务是当前移动通讯的一个增值业务，很多网站包括电视台的一些栏目，都支持短信与系统的互动交流。但互动交流都显得有很大的局限性，短信必须以预先设定好的格式来发送，否则电脑将无法识别。然而但是为了电脑处理的方便，这种服务局限性太大，在大多数的情况下，当我们想了解某种信息的时候，或者想发布某条信息到网站上的时候，我们更习惯于用自然语言的方式通过短信发送给系统，这就需要系统能够对自然语言进行识别，从中提取某些关键信息进行处理，从而理解用户的需求，跟用户进行交流。正是由于自然语言识别这一难题给用户和系统的交流带来了很大的障碍。

我们项目组在开发北京市失物招领系统的过程中，要实现通过手机短信的方式，把拾主捡到的物品信息通过手机短信的方式发送给我们的爱心平台，然后发布到网上。如果对于每一条用户发送的信息，都人工地进行分析，从中找出捡到物品的名称，捡到的时间以及捡到的地点，这将是一个非常繁重的任务，如何利用计算机快捷的计算速度来解决这一问题呢？这就只能寄希望于高效准确的自然语言识别了。

起初，我们用传统的模糊识别方法来解决这一问题，先把物品列表、地点列表放入数据库中，然后对输入的语言进行分

基金项目：国家 863 高技术研究发展计划资助项目（编号：2004AA115130）
作者简介：封春升（1981-），男，硕士研究生，主要研究方向：人工智能，虚拟现实，数据库应用。郝爱民（1968-），男，副教授，主要研究方向：虚拟现实，数据库应用。

析，检查输入的文本中有没有物品列表和地点列表中的内容，如果有则识别出来，没有就把它归为其他类的信息单独记录下来。在测试的过程中，我们发现这种方式经常识别错误。比如：

(1)地点识别不精确

例：今天我在王府井百货大楼捡到一个钱包。

模糊识别的结果是：地点为“王府井”，而正确的结果应该是“王府井百货大楼”。

(2)物品名称识别有歧义

例：今天，谁在车牌号为京 B2568 的出租车里丢了一把钥匙？

模糊识别的结果是：物品名称是“车”，而正确的结果应该是“钥匙”。

为什么会出现这种错误呢？这是因为我们在自然语言的识别过程中忽略了语义分析，为此我们提供了自己的一套算法，根据模式匹配来识别自然语言中的信息，这样可以大大增加识别的准确率。在语法分析的基础上，结合工程应用来定义最适合自然语言识别的语言模式。图 1 是识别的整个流程。

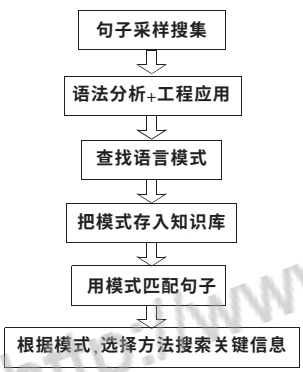

图 1　特定领域自然语言识别的流程图

2　研究对象的界定

由于我们的目的是实现一种高效准确的自然语言识别，所以为了集中精力解决问题的难点，我们特限定研究对象如下：

(1)对某一特定应用领域的自然语言识别。

(2)研究对象是单句或简单复合句。

3　关键技术分析

3.1　模式提取

为了用模式匹配的方法进行自然语言的处理，我们必须首先分析自然语言的语法、语义规则，并从中提取不同的语言模式，每一种模式都要有利于从中精确地检索出所需要的信息，这就是模式提取。

模式不同于语法语义规则的原因在于：模式提取不仅仅是分析语法语义，而是根据具体的需求，从自然语言的语法中提取某一成分，或者可以是几种成分的交融，作为模式，这样做的目的是更有利于根据模式识别出自然语言中所需要的信息，而不必受制于语法的限制。下面将针对于基于模式匹配的自然语言识别的方法，描述模式的形式化定义。

3.2　模式的形式化定义

(1)名词定义

准主语 Zsubject：是指语法成分中包含主语成分，但又可以附加主语成分周围的一些必要的文本信息。

准谓语 Zverb，准宾语 Zobject，准定语 Zattribute，准状语 Zadvmod，准补语 Zcomplement 的定义跟准主语的定义类似。

例：今天我在王府井百货大楼捡到一个钱包。

在定义模式的时候，为了能搜索出关键地名“王府井百货大楼”，我们定义准主语是“我在”，准谓语是“捡到”。

(2)模式的形式化定义

$P=\sum\{R,K\}\ [S]\ \{R\}<K>[S]\ \{R,K\}$

∑：表示其后的符号$\{R,K\}\ [S]\ \{R\}<K>[S]\ \{R,K\}$可以为一个也可以为好多个相联。

(3)符号说明

P：代表模式。

K：关键信息，即自然语言识别过程中所需要精确识别出来的信息。

S：模式要素，即模式定义过程中所需要的准主语、准谓语、准宾语等，是模式中最关键的部分。

R：冗余信息，即与所必需的关键信息及模式无关的其他信息。

{}：代表模式中可有可无的信息。

[]：代表模式中严格一致的信息，其中是语法成分 S。

<>：代表其中含有关键信息 K。

3.3　模式提取的原则

为了更好地从自然语言中识别出所需要的信息，首先要定义一个用来匹配句子的模式，模式的定义并不是一成不变的，这要根据不同的应用，不同的需求来规定。其定义的原则如下：

(1)由于几乎所有的自然语言都符合语法，所以我们根据汉语语法中的主、谓、宾、定、状、补来划分模式的每一个模块。

(2)找出所有需要识别的信息，并判断他们属于语法中的哪一成分(主、谓、宾、定、状、补中的某一个或某几个)。

(3)界定与关键信息有关的所有语法成分，把这些语法成分作为模式的必备模块。

注：这里的主、谓、宾、定、状、补并不是传统意义上的语法成分，而是以其为主，可以附加其他必要信息的准主语、准谓语、准宾语等。

3.4　模式分析

有了模式的形式化定义和模式提取的规则，对于具体的应用，我们又是如何来描述模式的呢？下面的分析都是以“要准确识别出自然语言中的地点和物品”为例进行模式分析的。

(1)把关键信息 K 周围的语法成分定义成模式要素 S。

模式 1　{R}[Zsubject]<地点>[Zverb]<R+物品>

为了识别出关键信息<地点>和<物品>，我们把准主语 Zsubject 定义为“我在”，而不是传统意义上的主语“我”，准谓语 Zverb 定义成为(“发现”，“拾到”，“捡到”)。

模式 2　{R}[Zsubject]<R+物品>[Zverb]<R+地点>

把准主语 Zsubject 定义为“谁把”，准谓语 Zverb 定义为(“丢在”，“落在”，“遗忘”，“忘在”)。

比较模式 1 和模式 2，很容易发现，由于关键信息<地点>和<物品>相对于模式要素 Zsubject 和 Zverb 的位置不同，所以把他们定义为不同的模式。我们这样做的目的是：

第一，能够大大提高关键信息<地点>识别的准确率，解决地点识别不精确的缺点。

例：今天我在王府井百货大楼捡到一个钱包。(用模式 1 匹配)

这样能够准确地识别出详细地点是“王府井百货大楼”,而不是“王府井”。

第二,能够限定关键信息<物品>的识别范围,提高识别的准确率。

还是以上面的那句话为例,我们把<物品>“钱包”的识别限定在准谓语“捡到”之后,这样可以缩小识别的范围。

(2)结合语义分析,而不仅仅从关键信息 K 和模式要素 S 的相对位置来提取模式。

模式 3 {R}[Zsubject]<地点>[Zverb]<R+物品>

把准主语 Zsubject 定义为“谁在”,而不是传统意义上的主语“谁”,准谓语 Zverb 定义成为(“丢失”,“落下”,“遗落”,“丢了”)。

例:谁在火车站丢失一个旅行包?

模式 3 和模式 1 虽然貌似相同,但因为其语义完全不同,一个是主动肯定,一个是被动疑问。通俗地说,也就是这两个模式的要素 S 不同,它们的准主语和准谓语是完全不同的,也不可能合在一起,所以不应该混为一个模式。

定义了这个模式之后,我们就可以解决背景介绍中提到的物品名称识别有歧义的问题了。

例:今天,谁在车牌号为京 B2568 的出租车里丢了一把钥匙?

根据定义的模式,我们把物品识别的范围限制在“一把钥匙”之内,所以就不可能错误地把物品识别为“车”了。

(3)对于复杂单句的分析,我们可以抛弃传统语法分析中某个语法成分只能出现一次的限制,根据只要适合于关键信息的识别,就可以把某个语法成分定义为模式要素的原则,来提取模式。

模式 4 {R}[Zverb1]<R+物品>[Zverb2]<R+地点>

这是一种复杂单句的模式定义,我们定义了两个准谓语 Zverb1 为(“发现”,“看到”),Zverb2 为(“丢在”,“在”),这样的定义更有利于关键信息的识别。

例:小王发现有个钱包丢在天安门广场。

3.5 模式匹配的级别设定

通过上面的分析,我们可以很好地对于特定的应用定义出完善的模式,模式定义完了之后,我们把它们放在知识库中,等进行自然语言识别的时候,我们先从知识库中提取模式,用模式来匹配自然语言,根据不同的模式,进行相应的识别。然而在模式匹配的过程中,我们应遵循什么样的原则呢?为了提高自然语言识别的速度,我们应该首先匹配最常用的模式,当匹配成功后就进行关键信息的识别,而不必再去匹配其他的模式,这就需要我们设定模式匹配的级别。级别的高低是根据统计规律获得的,在我们提取模式的过程中,我们已经获得了大量的实例,根据这些实例,我们可以判断出哪些模式出现的几率更大,出现的几率越大,模式级别就越高,就应该进行优先匹配。

4 基于模式匹配的自然语言识别的特点及应用趋势

基于模式匹配的自然语言识别,其最大的优点就是识别率高,识别速度快,并且自然语言识别系统简单,易构造,这就大大降低了自然语言识别系统的开发成本。它对于单句,以及复杂单句的识别非常准确,并且可以适当地推广到复合句中。但对于复合句的模式提取,复杂度将大大增加,我们还没有做这方面的研究,这还需要深一步的探讨。

基于模式匹配的自然语言识别方法对于任何一个特定领域都有着很广泛的应用价值,它能充分利用计算机的优势,自动地进行自然语言的分析和处理,大大地提高效率,缩小成本,必然会越来越多地应用于实际的系统中,并受到青睐,所以有着很广泛的应用前景。比如当前移动通讯广泛推广的 SMS 业务,对于短信发送和接收的信息,一般情况下都是语义比较简单的单句和复杂单句,对于这些信息的处理,完全可以用模式匹配的方法来实现对自然语言的处理。

5 实验系统与结论

为了论证基于模式匹配的自然语言识别的准确性和高效性。我们把这个识别系统应用于北京市失物招领系统当中的短信识别平台上(图 2)。

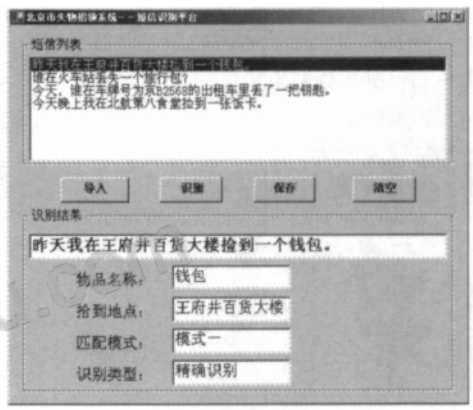

图 2 短信识别平台

由于北京市失物招领系统的短信业务目前还没有实现,为了验证自然语言识别的可行性,我们在网上发布了一个问卷调查,模拟接受用户丢失物品信息的发布,总共接收到 600 多条有效信息,并随机对其中的 100 条信息进行分析,按照上面论证的方法,进行模式的提取,并把所有的模式存入到知识库当中,用 C# 语言编写了一个基于模式匹配的识别算法,然后对所有的 600 条信息进行识别,正确地识别出其中 558 条信息的关键内容(物品名称及丢失地点),识别准确率高达 93%(图 3)。而且对于未识别出的信息可以记录下来,进一步分析其语法语义的特点,从而进行新的模式提取,并把这些新的模式存入知识库中,这样可以不断完善知识库,进一步提高自然语言识别的准确率。

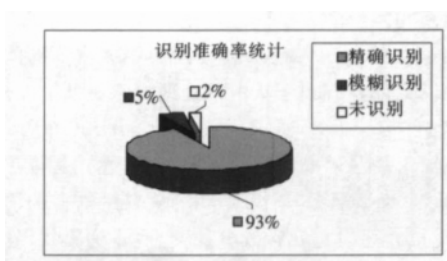

图 3 识别准确率统计

基于模式匹配的自然语言识别,打破了自然语言处理的传统方法,进行语义分析时,它不拘泥于传统的语法成分,而是根据需求,灵活且适当地扩展了语法成分的含义,更有利于关键信息的识别,而这其中最核心的工作就是进行模式提取。基于

(下转 166 页)

步骤 3 对于测试集 S' 中的每个样本 x_i：

(1)用 $NB^{(N)}$ 对 x_i 分类。

(2)将 x_i 加入到 T' 中并初始化其权值为 $\mathrm{Min}w_i(1\leqslant i\leqslant M)$。

(3)$M=M+1$。

(4)将 x_i 从 S' 中去除。

(5)转向步骤 2。

输出：分类器 NB。

4 实验及分析

本文以 AdaBoost 提升方法为参照进行分析，在 UCI 机器学习数据库中的 18 个数据集上进行了实验。由于 ActiveBoost 仅能处理离散数据，我们采用基于信息熵[8]的全局离散化方法对连续属性进行预处理。在有丢失值的数据集中，以出现频率最高的值填充。

将数据集 DS 通过随机选取的方法分为训练集 LS 和测试集 TS，并令 $TS=DS\times 10\%$。经过 10 倍交叉验证估计分类器的正确率。在每个数据集上分别测试 20 次，每次实验采用不同的 10 重划分。表 1 列出了 20 次测试的平均正确率及标准偏差，采用双边 t 检验比较这两种提升方法。“√”在显著性水平 0.05 下 ActiveBoost 优于 AdaBoost。

表 1 预测准确度和基本偏差比较

Data set	AdaBoost	ActiveBoost
Anneal	98.312 5±1.528 4	98.637 3±1.827 3√
Audiology	77.304 4±7.528 4	82.156 5±6.937 5√
Australian	85.625 7±4.073 3	83.964 3±4.834 5
Breast-w	93.672 6±2.674 3	94.784 6±2.943 5√
German	71.338 1±3.294 3	75.837 4±2.845 7√
Glass	67.662 1±9.332 1	65.473 2±8.703 8
Heart-c	76.931 5±6.916 2	83.743 9±6.492 3√
Heart-h	80.272 5±8.018 3	85.857 4±6.860 4√
Ionosphere	89.783 6±4.493 2	88.875 0±3.967 7
Iris	94.716 2±5.394 2	96.648 5±4.764 4√
Kr-vs-Kp	99.403 5±0.492 7	97.634 3±0.085 4
Pima-indians	74.512 0±5.363 7	77.754 3±4.083 3√
Primary-tumor	41.415 8±6.921 7	46.632 2±6.548 0√
Segment	96.893 7±1.603 8	95.863 4±1.085 3
Sick-enthyroid	98.789 2±0.694 8	96.975 3±0.978 3
Soybean	91.882 4±3.294 8	93.864 5±2.075 3√
Vehicle	72.394 3±4.203 6	80.755 3±3.765 4√
Zoo	92.646 2±7.649 4	94.655 4±6.653 3√

从表 1 中可以看出，ActiveBoost 在 12 个数据集中对朴素贝叶斯预测准确度的改善程度优于 AdaBoost。说明本文提出的主动提升策略是有效的。而更令人感兴趣的是，基于主动学习可以挖掘出未分配类别标注的样本中的信息。这在较难获得样本类别标注的场合下是非常有用的。

5 结论

本文结合主动学习将不确定性引入到朴素贝叶斯的构造过程，同时挖掘未分配类别标注的样本中的信息，在基准数据集上的实验结果表明，本文提出的 ActiveBoost 表现出较优的分类性能。(收稿日期：2005 年 11 月)

参考文献

1.Schapire R E.The strength of learnability[J].Machine learning,1990;5(2):197~227

2.Freund Y,Schapire R E.A decision theoretic genelization of online learning and an application to boosting[J].Journal of Computer and System Science,1997;55(1):119~139

3.Schapire R E,Singer Y.Improved boosting algorithms using confidence related predictions[C].In:Proceeding of the 11th Annual Conference on Computational Learning Theory,1998:80~91

4.E Bauer,R Kohavi.An empirical comparison of voting classification algorithm:Bagging,boosting and variants.Machine Learning,1999:105~142

5.K M Ting,Z Zheng.Improving the performance of boosting for naive Bayesian classification[C].In:Proceedings of the 3th Pacific-Asia Conference on Knowledge Discovery and Data Mining,Berlin Germany:Springer-Verlag,1999:296~312

6.Yonglong W et al.MIQR Active Learning on a Continuous Function and a Discontinuous Function[J].Neural Computing and Applications,2001;10(3):253~270

7.Hong L Shang-teng H.A Genetic Semi-supervised Fuzzy Clustering Approach to Text Classification[C].In:Proceedings of the 4th International Conference on Web-Age Information Management,Chengdu,China,2003:173~180

8.Dougherty J,Kohavi R,Sahami M.Supervised and unsupervised discretization of continuous features[C].In:Proceedings of the 12th International Conference on Machine Learning,Morgan Kaufmann Publishers,San Francisco,CA,1995:194~202

(上接 146 页)

模式匹配的自然语言识别吸取了基于关键字匹配的方法和基于大规模语料库的自然语言处理方法的优点，提出了一个折衷的方案，更加适合于工程应用，而不仅仅是理论分析。由于其开发成本相对较低，和识别准确率较高这两个显著的优点，所以对于自然语言识别走向实际应用提供了一个很好的解决方案。(收稿日期：2006 年 1 月)

参考文献

1.Hughes,John.Automatically caquiring a classification of words[D].Ph D Thesis.University of Leeds,Paris,1994

2.Zavrel,Jakub.Lexical Space:learning and using continuous linguistic representations[D].Ph D Thesis.Tilburg University,Tilburg,1996

3.周雪忠，吴朝晖.文本知识发现：基于信息抽取的文本挖掘[J].计算机科学，2003;30(1):63~66

4.龚小谨，罗振声，骆卫华.汉语句子谓语中心词的自动识别[J].中文信息学报，2003;17(2):7~12

5.鲁松，宋柔.汉英机器翻译中描述型复句的关系识别与处理[J].软件学报，2001;(1)

6.绍军力，张景，魏长华.人工智能基础[M].北京：电子工业出版社，2000