基于决策树方法的口语句子边界切分

周云 黄泰翼
中科院自动化研究所
模式识别国家重点实验室 北京 100080

摘要： 本文提出了一种基于决策树的并结合规则的新方法来解决口语句子边界自动切分的问题。通过与基于 N-gram 的统计方法比较，结果是令人满意的。精确率达 82%,召回率达 75%。最后分析了切分产生错误的类型和主要原因及有待进一步研究的问题。

关键词： 决策树，基于规则的解析，口语句子边界切分，

Utterance Segmentation Based on Decision Tree Method

Yun Zhou and Taiyi Huang
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing
{zhouyun, huang}@nlpr.ia.ac.cn

Abstract :In this paper we propose a decision tree-based method, which combined with rules for utterance segmentation in spontaneous speech. Compared with N-gram method, the results are satisfying: The performance of precision 82% and recall 75% are reached by proposed approach. In the end of this paper we analyze the causes and type of errors and put forward some issues to be studied in the future.

Keywords: decision tree, utterance segmentation, rule-based parsing

1. 前言

对话口语与书面语有着许多不同的特点[1],其中一个显著的区别是口语对话中的一轮是没有边界的。因为口语不象书面语那样有标点符号。由于许多对话计算模型是将口语句子而不是一轮作为基本处理单元，因此，口语边界切分这一问题十分重要。另外，在对话口语翻译系统中识别器的输出是按声学层次如停顿，静音等进行切分的。而翻译器常常需要以语言学层次上的口语句子作为基本输入单元。在这些情况下对口语边界的切分也是十分必要的。尽管在人机对话中似乎绝大部分情况都是一轮只包括一个口语句子，但在人人对话中一轮包括多个口语句子的情况是常见的。

口语句子切分中可以用到的特征大致分为以下 3 类：（1）韵律特征，包括基频，重音，时长，停顿等[2]（2）线索词[4]，即某些标志口语句子的起始或结束的词（3）词,词性，语

义[3][4][5]。用到的典型算法有神经网络[2][5]，基于统计的 N-gram 模型[3][4]等。
本文提出了一种基于决策树并结合规则的口语边界切分算法。决策树学习算法是机器学习中的一类重要并且得到广泛应用的方法，它属于归纳学习，旨在从大量的经验数据中归纳抽取一般的判定规则和模式。决策树学习的方法适合解决带有如下特点的问题：第一，例子能够用“属性－值”这样的模式对来表示；第二，目标函数具有离散输出值；第三，允许训练数据有错误；第四，训练数据可以有某些属性的缺少。第五，通过扩充问题集，决策树方法易于添加新属性。从下文可以看出决策树方法的这些特点非常适合于解决口语对话语句的自动切分问题。本文中决策树所用的属性为词的词性、语义和是否属于线索词，没有用到具体词的信息，这主要是考虑到算法的时间复杂度的问题。为部分弥补这一缺陷，本文采取了添加规则的方法来处理某些特定词的断句情况。通过与文献[3]中的 N-gram 统计方法比较，结果是令人满意的。本文提出的方法的精确率达 82%，召回率达 75%。

2. 对话语料的预处理

我们进行训练和测试的语料是有关旅馆房间预订的录音。共 94 段，为两人对话方式。语料由 16254 个词，3008 轮组成，并由专人将录音材料抄写成文本形式。所有语料都是自然状态下的录音，语料中包含了丰富的口语语言现象，具有代表性和真实性。
对所有的词我们确定了 18 个词性类[6]55 个语义类[7]和 4 个线索词。其中线索词共 4 个，分别为：是这样，就是说，这样，那这样。语义类是对 18 类词性中某些类词的细分。如我们将名词细分为房间类，价格类，单位类，设施类等。词性和线索词同领域无关，而语义类则是和领域密切相关的。我们对语料进行了词性、语义和线索词的标注，这些都是决策树算法将提取的属性特征。考虑到算法的可移植性问题，移植到新领域时，只需重新设计与领域有关的语义类，而词性类与线索词都与领域无关，无须变动。
什么是口语句子，口语句子的语言学上的切分标准是什么，这些都是语言学界有争议的问题。我们从面向特定领域的口语理解这一实际目标出发，从语用,语义和语法角度制定了 6 条汉语对话口语断句标准，请参见附录。用这些标准我们对全部语料 3008 轮进行切分，得到共 4875 个口语句子[1]。平均每轮含有 1.62 个口语句子。以下是一个切分实例：

a:您好/我要订一个礼拜天中午的单间
b:单人间/告诉我客人的名字

3.算法实现

3．1 基于词的信息提取

假定一轮中有 n 个词 $W_1\ W_2 \ldots\ W_n$，前后补上两个空位置得 $\Phi\Phi W_1\ W_2 \ldots\ W_n \Phi\Phi$，对每一个非空位置 i($1 \leqslant i \leqslant n$)的词，取该位置 i 前后各两个词共五个词组成窗口 $W_{i-2}\ W_{i-1}\ W_i\ W_{i+1}\ W_{i+2}$ 在训练语料中，每个非空位置 i($1 \leqslant i \leqslant n$) 所提取的信息 $Info\text{-}Train_i$ 是一个 8 元式：

[1] 有的文献中将每个口语句子称为一个 utterance

Info-Train$_i$＝{P_{i-2} ,P_{i-1}, P_i ,P_{i+1}, P_{i+2}, Sem_{i-1}, Sem_i ,Cue_i, R_i },各项含义如下：

$$P_i=\begin{cases}p, p是位置为i的词的词性, i\in[1,n], p\in[1,18]\\ Null, \qquad 否则\end{cases}$$

$$Sem_i=\begin{cases}s, p是位置为i的词的语义类, i\in[1,n], s\in[1,55]\\ Null, \qquad 否则\end{cases}$$

$$Cue_i=\begin{cases}c, 位置为i的词属于第c类线索词, i\in[1,n], c\in[1,4]\\ Null, \qquad 否则\end{cases}$$

$$R_i=\begin{cases}T, 位置为i的词是句首, i\in[1,n]\\ F, \qquad 否则\end{cases}$$

在测试语料中每个非空位置 i(1≤i≤n) 所提取的信息 Info-Test$_i$ 是一个 7 元式：Info-Test$_i$＝{P_{i-2} ,P_{i-1}, P_i ,P_{i+1}, P_{i+2}, Sem_{i-1}, Sem_i ,Cue_i}, 与 Info-Train$_i$ 的区别仅在于没有 R_i, 其余各项含义与 Info-Train$_i$ 相同。口语对话切分问题即为对每一个位置 i 的 Info-Test$_i$ 找到与其对应的 R_i 也就是建立{Info-Test$_i$}→{T,F}的映射.T 代表位置 i 是口语句子起始切分点，F 代表不是句子起始切分点。

以下举一训练语料中的实例说明信息提取过程：

a: 您好(I)/单人间(N) 有(V)吗(Y)

位置 1 为词“您好”，词性为惯用语(I),对应词性的第 6 类表示，即用数字 6 表示惯用语。但“您好”不是线索词，所以其线索词属性为空，“您好”也不属于我们定义的 55 个语义类中的任何一个，所以它的语义类属性也为空。”“您好”为口语句子起始点，定为“ T”。依此类推，得如下表格形式：

表 1 训练集基于词的信息提取

位置	属性	P_{i-2}	P_{i-1}	P_i	P_{i+1}	P_{i+2}	Sem_{i-1}	Sem_i	Cue_i	R_i
1		Φ	Φ	6	1	2	Φ	Φ	Φ	T
2		Φ	6	1	2	9	Φ	24	Φ	T
3		6	1	2	9	Φ	24	37	Φ	F
4		1	2	9	Φ	Φ	37	Φ	Φ	F

每一个位置连同它的属性我们称为一个例子。所有训练语料中的例子的集合称为训练集。当 R_i 取 T 时我们称为正例，当 R_i 取 F 时我们称为反例。所有的正例组成正例集 T-set，所有的反例组成反例集 F-set。在全部 94 段对话语料中我们取 84 段作为训练语料，其余 10 段作为测试语料。T-set 有正例 4065 个，F-set 有 10541 个。

3．2 决策树中问题集的设计

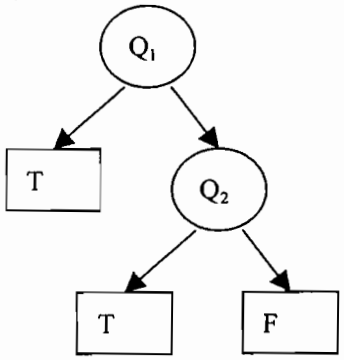
为了问题集的易扩充性及决策树能够用二叉树形式表示，我们以如下的方式设计问题：

对 Info-Train 中的每一个属性,我们在其取值范围内的是非提问。如对当前位置的词性 P_i 提如下 V 个问题（V 代表词性总数，此处为 18）：是否是第 1 类词。。。是否是第 V 类词。对这 V 个问题的回答必然是其中一个为真，其余 V-1 个为假。这样我们共得到 18×5＋55×2＋4＝204 个问题，形成问题集 $Q=\{q_1 q_2 \ldots q_i\}$,$q_i \in \{1, 0\}$,其中，1 为真，0 为假，$1 \leqslant i \leqslant 204$。

3．3 决策树训练算法

我们采取决策树算法中比较常用的一类算法－ID3 算法[9]。整个训练算法返回一棵能正确对训练集进行分类的决策树

图 1 中给出了一个算法生成的二叉决策树的示意图。



注 （1）图中圆形代表非叶节点，矩形代表叶节点
（2）圆中 Q_1Q_2 的代表该节点从问题集中选择的问题
矩形中 T, F 代表最终分类结果－为切分点或不为切分点
（3）左子树代表对问题的回答为“是”，右子树对应回答为“非”

图 1　二叉决策树示意图

3．4 规则的添加

考虑到时间复杂度问题，本文采用的决策树算法只用到词的语法、语义、以及是否是否线索词这 3 类信息，词的其他特征如搭配特征等没有用到。另一方面，由于训练语料规模有限，为了提高决策树的归纳能力。所以我们要对决策树算法的切分结果再进行基于规则的处理。

我们制定的规则采取如下的形式：Conditions=>Action。Conditions 为规则的前提条件，一般由关键词或词性语义等表示。Action 为满足此条件时将当前位置设为切分起始点或非切分起始点。对决策树的切分结果，若满足某条规则的前提条件，就执行这条规则。

我们的规则分为两类：（1）对决策树错误切分结果的修改。如：IF（（当前词＝“下”）&&（前一词＝“一”））THEN（当前词位置不切分为起始点）。（2）对特定词的处理规则。如：IF（（当前词＝“那”）&&(当前词词性＝连词)）THEN（当前词位置切分为起始点）。

4.实验结果及分析

4．1 实验结果

本文的切分算法对每一轮中每个位置的词做出是否为句首切分点的判断。我们采用精确率 precision 和召回率 recall 两个主要性能指标来评价算法性能，定义如下

$$precision = \frac{M}{N}, \quad recall = \frac{M}{H}$$

其中 M 为被算法正确判定为句首切分点的个数，N 为所有被算法判定为句首切分点的个数，H 为测试语料中所有应该被切分为句首切分点的个数。

精确率与召回率是系统都要考虑的评价参数。为了全面衡量系统的性能，我们另外还采用了 F一测试[10]。定义如下：

$$F = \frac{(\beta^2 + 1) \times prcision \times recall}{\beta^2 \times precision + recall}$$

其中 β 为平衡精确率与召回率的一个系数，此处我们取 β=0.5，1 和 2 三种情况。94 段语料中我们取 10 段共 266 轮作为测试集，另外 84 段共 2742 轮作为训练集。

实验结果如下：

表 2 实验结果

	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6
精确率	61.2%	75.0%	76.1%	81.9%	55.2%	68.1%
召回率	58.2%	69.6%	70.4%	74.8%	64.0%	73.8%
F一测试 β=0.5	60.6%	73.9%	74.7%	80.3%	56.8%	69.2%
F一测试 β=1.0	59.7%	72.2%	73.0%	78.1%	59.3%	70.8%
F一测试 β=2.0	58.8%	70.6%	71.4%	76.1%	62.0%	72.6%

上表中，模型 1，2，3 为只用决策树算法。模型 1 用的特征是词性。模型 2 用的特征是词性加语义。模型 3 用的特征是词性、语义以及线索词。模型 4 是在模型 3 的基础上添加规则。模型 5，6 采用的是文献[3]中用到的 N-gram 统计模型。模型 5，6 都采用 Tri-gram，并用 Vertibi 算法求出一轮中的切分点。模型 5 所用特征为词性，符号集为 18 个。模型 6 用的特征为词性和 23 个关键词[2]，符号集为 41 个。所有模型的训练集和测试集和测试集都是相同的。

对比模型 1，2，3 可见语义特征的加入对断句的准确性有较明显的帮助。对比模型 4，6，可以看到，我们提出的决策树加规则并引入了语义特征的方法，对词进行了更精细的刻画，总体性能优于基于 N-gram 统计模型的方法。对比模型 1，5，两者均只用到词性特征，决策树与 N-gram 两种方法处于相近的水平上。这也从另一方面说明断句问题中细

[2]连词、语气词及线索词等

致的语言学特征对提高系统总体性能效果更明显一些。

另外，为了研究扩大观察窗口对分析性能的影响，我们在模型 4 的基础上将观察窗口长度扩大到 7，即以当前词为基准，取前 3 词和后 3 词的词性，其余不变。但得到的精确率和召回率与模型 4 没有改变，仍然分别为 81.9%和 74.8%。这说明远距离的上下文信息对当前词的断句决策没有或很少有影响。

4．2 切分错误分析

算法标注结果产生的切分错误可分为两类：

（一）误切分

即不该断的地方断开了。从误切分产生的位置来看，又可分为两种：

（1） 误切分点在基本短语边界

这种情况占误切分的 72.8% 如：

例 1： A：另外呢/房间的话

例 2：A：您先按他的名字/登记两个单人间

（2） 误切分点在基本短语内部

这种情况占误切分的 22.2% 如：

例 3：A: 明天/上午到吗

产生误切分的原因有的是因为口语现象造成的，如例 1；有的是因为提取特征的窗口长度有限，一轮内部缺少更长距离的结构信息，如例 2；

(二) 漏切分

即该断的地方没有断开。如：

例 4：A：标准间还有吗

B：有您什么时候要（正确切分为：有/您什么时候要）

例 5：A：单人间有一百三（正确切分为: 单人间有/一百三）

例 6：A: 要不我过一会儿打过来您先查一下（正确切分为: 要不我过一会儿打 来/您先查一下）

产生漏切分的原因有的是因为从局部的一轮很难准确断句，需要结合上下文语境分析，如例 4；有的是因为对话者高度省略所造成，如例 5；有的是因为切分点缺乏明显断句特征，如例 6。

5.结束语

本文提出的对话口语边界自动切分算法采取了决策树与规则相结合的方法。。。所提出方法的精确率达 82%，召回率达 75%，优于与文献[3]中的 N-gram 统计方法，结果是令人满意的。

但是也应该看到，口语句子切分是一个十分复杂的问题。韵律信息的缺少使得我们的算法无法达到很高的精确率和召回率。如何准确提取韵律这一声学层特征，并与词性语义等语言学层次的特征有效地结合来一起解决口语句子切分问题，是今后值得研究的一个问题。另外，我们的算法要求正确输入每个词的词性。这里我们假定词性完全正确输入，但在实际系统中词性标注不可能完全正确。词性标注的正确与否对切分的影响也有待进

一步研究。

参考文献

[1] Yun ZHOU, Taiyi HUANG, Bing ZHAO, *The Analysis of Corpus Oriented to Spoken ChineseDialogue Understanding* , International Symposium of Chinese Spoken Language Processing -2000(ISCSLP2000) Oct. 2000

[2] Warnke et al. ,*Integrated dialog act segmentation and classification using prosodic features and language models* ,Eurospeech-97, 1997,pp 207-210

[3] Stolcke,A and Shriberg,E., *Automatic linguistic segmentation of conversational speech*， ICSLP-96, 1996,pp 1005-1008

[4] Heeman,P.A. and Allen,J, *Speech repair, intonational phrases and discourse marker: Modeling speakers' utterances in spoken dialog*， 1994,Computational Linguistics 25(4)

[5] 王海峰等, *汉语口语语段边界的界定* ,计算机学报, 第 22 卷 第 10 期，1999

[6] 宗成庆, *口语自动翻译方法研究*， 中科院模式识别实验室博士后出站报告,附录 A，2000

[7] 邓云滨,*口语对话系统领域移植---统计语言理解*，中科院模式识别实验室硕士论文，2000

[8] 洪家荣, *归纳学习—算法 理论 应用* ,科学出版社, 1997 pp33-37

[9] Tom. Mitchell , *Machine Learning* ， MIT Press 1997 , pp57-58

[10] Daniel et al , *Speech and Language Processing*, Prentice Hall 2000 pp 578

附录 断句规则

我们制定的 6 条断句规则如下：

1 划分的每一个句子，除非属于无法理解或不完整的句子，都要与一个且只有一个 Speech Act 对应

2 充当一个句子语法成分的部分不能分开

3 所有 speech repair 现象都在一个句子内部

4 对无法理解的句子，不完整句子单独划开

5 对一个句子主干成分的补充说明，除非有谓词引导，否则不能单独成句子

6 如果句子中充当谓语的是动词或动词词组，则所有从属于谓语中心动词语义框架的成分，无论是必要的还是可选的，都必须在一个句子内。