

限定领域汉语口语对话语料分析

宗成庆 吴华 黄泰翼 徐波

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 口语对话语料分析是口语交互翻译系统和人机对话系统研究的基础。本文以限定领域汉语口语对话语料为基础，首次对汉语口语对话中的语言现象进行了详细统计和分析，提出了建立通用口语词典和适应不同应用领域转移性的词汇提取方法。本文提出的方法和统计的结果对于研究鲁棒性汉语口语理解算法和建立限定领域的口语翻译系统和人机对话系统，具有重要的意义。

关键词 话语分析 语料统计 口语翻译 人机对话

Analysis of Spoken Dialog Corpus in Restricted Domain

ZONG Chengqing, HUANG Taiyi and XU Bo

cqzong@nlpr.ia.ac.cn

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100080

Abstract Analysis of spoken dialog corpus is an important foundation for research of speech-to-speech translation system and man-machine dialog system. Based on the Chinese spoken dialog corpus in restricted domain, this paper firstly presents statistical data and performs analysis in detail on the spoken language phenomena. The ideas to establish universal spoken language dictionary and the methods to extract words from corpus of different domains for expanding dictionary are also presented. The methods and statistical results presented in the paper are meaningful for research of robust Chinese spoken language understanding and implementation of spoken language translation system and man-machine dialog system.

Keywords Discourse analysis, Corpus statistics, Spoken language translation, Man-machine dialog system

1. 引言

在口语翻译和人机对话系统研究中，口语语料的收集、整理和分析是非常重要的基础性工作。由于口语对话中的语言现象与书面语相比具有较大的差异，而且不同的应用领域

之间，词汇的分布等各种语言现象也有所不同，因此，如何建立有效的语料分析方法，对特定领域的口语现象进行准确地统计和分析，对于口语理解方法的研究和建立限定领域的口语翻译系统和人机对话系统，具有重要的意义。

近几年来，随着口语翻译系统和人机对话系统的研究日渐兴起，口语语料的分析受到越来越多的重视，尤其在国外，话语分析(Discourse Analysis)已经成为自然语言处理研究中的一个比较活跃的分支^[1-3]。而在我国，除了语言学家对汉语口语进行过一些定性的分析^[4]以外，在自然语言处理领域中对口语现象的研究还只是十分初步的。本文以旅馆预订领域收集到的对话语料为基础，对汉语口语中的语言现象进行了详细的统计和分析，提出了建立通用口语词典和适应不同应用领域转移性的词汇提取方法。本文统计的结果和提出的方法，对于研究鲁棒性汉语口语理解算法和建立限定领域的人机对话系统及口语翻译系统，具有较大的参考价值。

2. 语料收集及统计结果

本文的研究工作以旅馆预订领域的对话语料为基础。对话过程是两个人通过电话以一问一答的形式进行的，电话一端代表客户，另一端是旅馆前台服务人员，说话方式完全是自由的和随意的。

为了表述方便，我们首先给出如下定义：

定义1 对话语句(Utterance) 从对话者一方开始讲话到讲完停下或被对方强行打断为止，所说的全部内容称作一个对话语句。

定义2 对话子句(Dialog sentence) 一个对话语句中所包含的分句，称作对话子句。

例如：我打电话到订票处了 / 他说票特别紧张 / 他说去试一试 / 这样吧 我明天一早 给您挂电话 行吗

这一段文字从开始到结束是一个对话语句，在这个对话语句中包含有4个对话子句（由“/”隔开）。

2.1 语料的整理与标注方法

口语对话语料的收集与书面语语料的收集方法不同，它首先需要通过录音的方式将对话内容记录下来，然后再将其整理成文字，因此，往往需要大量的手工操作。首先用录音电话将对话内容记录在磁带上，然后由人根据磁带记录的信息将对话整理成文字。整个处

理过程可以简单地表示为如下流程图：

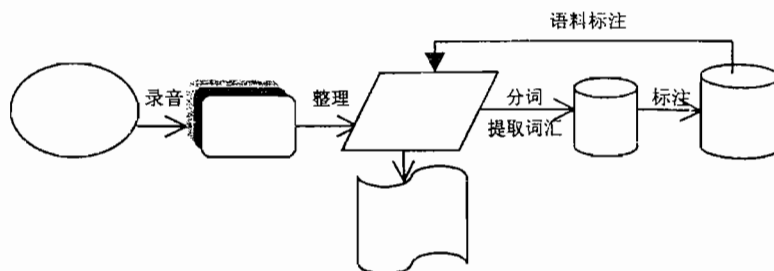


图 1 语料整理流程

其中，词典标注采用人工方式，语料标注采用机器自动标注与人工核对相结合的方法。当第一次收集语料时，词典是不存在的，我们只需要对整理的语料进行分词，然后提取所有的词汇，并依据这些词汇建立相应的词典。词典一旦建立之后，当领域转移或扩展时，系统只需将新收集的语料中在词典中没有的所有新词提取出来，然后在人的辅助下决定是否将这些新词添加到词典中去。这样，对于应用领域接近或同一领域扩展时，只需要处理少数新的词汇即可，而避免了大量的重复工作。

2.2 词类划分与词典标注

面向口语翻译系统，我们将建立一个通用的汉语口语词典。词典中尽量收集口语中使用较多的一些口语词汇和相对稳定的一些常用的虚词。每一个词条信息包括汉语词条的词类、语义特征信息和对应的英语单词等。

考虑到口语中若干词的使用方法和含义与其在书面语中的作用和含义有所差异，因此，在考虑一般的汉语词类划分方法的同时，我们还专门针对口语中的词汇进行了调整，将全部词汇划分为18个大类：名词N、代词P、时间词T、处所词W、动词V、助动词X、判断动词J、形容词A、数词Q、副词D、方位词F、介词R、连词C、助词H、量词L、语气词M、拟声词Y、习惯用语I。其中，习惯用语较多地考虑了日常口语对话中的用词习惯，它包括敬语、插入语、感叹词和呼应语以及搪塞性的词语。呼应性的词是指对话双方在交互过程中应答性的口头用语，如：“嗯，好的，哦，没错儿，够戗，”等；敬语主要包括日常对话中的一些寒暄性用语，如：“多谢，不敢当，久仰”等；搪塞性的词语是指当说话者处于思考状态或语塞状态时使用的一些，如“那个、这个、这”等。

词义特征标记根据不同的词类采用不同深度的树状层次化结构的标记方法，由上位到下位逐步细化，层次最多的是名词，语义特征树的高度最高达8层。

2.3 语料统计结果

根据上述介绍，我们对收集到的100多段涉及旅馆预订的对话语料进行了全面统计。

以下给出部分统计结果：

(1) 词长分布

口语中的词长分布如下表所示：

表 1 词长分布

词长(字)	1	2	3	4
比例(%)	28.51	57.20	12.99	1.31

统计结果显示，在口语对话中，1字词和2字词占绝大多数(86.19%)，3字词和4字词只占少数，4字以上词基本上极少出现。口语的平均词长为1.87个汉字，比书面语的平均词长(约2.45汉字)短^[5]。

(2) 对话语句长度分布

对话语句长度的分布情况如下：

表 2 语句长度分布

长度(字)	1	2	3	4	5	6	7	8	9	10	11-67
比例(%)	15.12	8.34	9.28	8.54	7.68	6.78	5.27	5.27	4.78	4.09	24.85

在我们收集到的语料中，最长的对话语句为67个汉字，占全部语句的0.08%，其次长度是61个汉字，占0.041%。长度为1的语句数最多，这些语句一般是单字长的语气词或呼应性的单字词，如：啊，噢，嗯等，平均语句长度为7.8个汉字。

(3) 词类分布统计

根据我们对口语词类的划分标准和对口语词的处理方法，对收集的语料进行统计后得到如下词类的分布结果：

表 3 词类分布统计结果

词类	A	C	D	F	H	I	J	L	M
比例(%)	4.00	1.52	6.84	0.52	3.98	10.77	2.63	2.87	5.37
词类	N	P	Q	R	T	V	W	X	Y
比例(%)	14.69	10.88	15.61	0.66	3.10	15.31	0.47	1.63	0.00

从统计结果我们可以看出，口语中使用最多的5种词类依次是：数词、动词、名词、代词和习惯用语。在统计语料中数词之所以如此之多，主要是因为是在旅馆预订时经常需要交换电话号码、询问费用和房间号码、楼层号等原因所致。

(4) 非规范语言现象的统计

我们专门对口语中经常出现的重复、次序颠倒、冗余和语句残缺（严重省略）等4种非规范语言现象的出现几率进行了统计。以下首先说明4种非规范语言现象的界定：

1) 重复

这里所说的重复主要指字面上明显的重复现象。如下面的例句：

(a) 啊 打九折行 下礼拜下礼拜二三吧 好吗

(b) 不不 现在先不订 过两天再订

2) 次序颠倒

次序颠倒主要指语序上明显的不合理现象。如下列例句：

(c) 已经给您打过折扣了 这个

(d) 有房吗 现在

3) 冗余

冗余现象主要指语句中含有多余的词语。如下面的例句：

(e) 那个可以预订吗 可以

(f) 就里面的那些设施啊怎么样 就是条件 我想大概就是说看一下

4) 省略或语句残缺

省略现象主要指说话人根据对话主题和上下文语境，有意省略双方正在讨论的问题所对应的词汇，或者说话者出现语塞，或被对方强行打断等现象引起的对话语句残缺。如下面的语句：

(g) 那六楼呢

(h) 差别在哪里 二百三百

另外还有一类语句我们把它们称之为独词句或零句，这类语句一般只有一个词或同一个词连续重复，多数是说话一方呼应对方或彼此使用的敬语，如：“对、对、对”，“不客气”，“没问题”等。独词句和省略句只按一种情况记。

以下是对上述4种非规范语言现象和独词句出现几率的统计情况：

表 4 非规范语言现象出现几率

语言现象	重复	次序颠倒	冗余	省略	独词句	现象并存
出现几率 (%)	3.56	1.23	4.70	32.61	44.59	5.68

其中，“现象并存”是指重复、次序颠倒、冗余和省略（或独词句）至少两种语言现象同时存在的情况。

另外，口语中常常因为说话人边思考，边重复对方讲话内容，同时还伴随介绍正在进行的动作，从而导致说出来的语句支离破碎。如：

(i) 住到 我看看啊 住两天两晚上

(j) 单人间 告诉我客人的名字

系统要对类似的句子正确地理解，必须能够识别并分隔这些语句中的插入片段或子句。

3. 口语歧义现象分析

从上面的统计结果我们可以看出，在口语会话中无论是用词情况，还是句子结构等各方面，与书面语相比都有较大的差异。除了上述已经获得的统计结果以外，口语中出现的歧义现象也十分复杂，以下我们对口语中的歧义现象进行分析。

3.1 口语中的词汇歧义

同书面语中的词汇歧义一样，口语中的词汇歧义主要是由词汇的一词多义引起的。在口语词类的划分中，尽管有些词按严格的汉语词法规定不能算作词，如“不好、不行、行了”等，但考虑到它们在口语中大量地存在，并且含义相对固定，因此，我们将其作为一个词处理。但是，这些词在不同的场合下使用，有时仅仅是因为语气的差异而使其意思完全不一样。请看如下例子：

(1) a. 这样肯定不行。

b. 不行您过来一趟。

在例句a中，“不行”表示对结果的一种判断，是确定性的；而在例句b中，“不行”仅仅表示一种假设的语气，相当于“要不然的话”或“要不”。

(2) a. 您说的那个宾馆不是我们这里。

b. 价钱我们不是已经说好了吗？

第一句中的“不是”是否定意义的断词，在句子中充当谓语成分；而第二句中的“不是”仅作为强调语气的一个填充性的词语，本身在句子中并不充当什么成分，含义也没有任何否定的意义，相反加强了肯定的语气。

类似地，“不好，什么，那个”等在口语中都有不同的语义表示。

3.2 口语中的结构歧义

由于口语语句中没有标点符号，又缺少了必要的声调、语气和停顿信息，因此而引起

语句的歧义结构也是常见的现象。如下例子：

(1) 您想住双间不在一层可以吗

这个句子由于丢失了语气（声调）信息而可能引起歧义理解。它可以有如下两种理解结果：

(a) 您想住双间不？在一层可以吗？

(b) 您想住双间，不在一层可以吗？

结果(a)将例句(1)拆分成两个子句，问的是两个问题；结果(b)为一子句，前面部分好象是讲话人在重复对方的要求，后半部分才是要问的问题。而“一层”在(a)、(b)中也可以分别有两种不同的理解，(a)中一般理解为“第一层”（first floor），而在(b)中一般指的是“（与别人）同一层”。

(2) 如果是二月五号二月四号走也行

该句至少有如下两种不同的理解结果：

(a) 如果是，二月五号二月四号走也行。

(b) 如果是二月五号，二月四号走也行。

两种理解结果差距甚远。

3.3 口语中的指代歧义

口语中的指代歧义主要指对话双方由于过多地使用省略或频繁使用口语中的代词而引起的理解误导。如下面的对话：

A: 你好 我是XX公司的 我姓张 张XX 常到您那去 咱是老朋友了

B: 唉 您好 张先生

A: 最近咱公司要来两个客人 准备在咱们这住 我问一下咱这能打折不

... ..

在本段对话中，说话者A方分别在第一句和第三句多次使用了“咱”一词，有时表示的是己方，有时又表示对方，对话双方由于特定的场景，可能会理解其中的含义，但是，对于分析器来说，却难以分清各自所指。

4. 结论与结束语

根据上述统计结果和分析的实例，我们可以得出如下结论：

- 口语中的词义相对简单。口语中的平均词长比书面语中的平均词长要短，每个词

所表达的义项也相对简单，尤其在特定领域中，每个词几乎就只有一个含义。这样对于口语翻译系统或对话系统来说，词法分析的任务就相对简单了。

- 口语语句结构相对简单。在我们收集的语料中，口语语句的平均语句长度只有7.8个汉字，语句结构基本都是简单句，尤其独词句占有相当的比例，这对于句法分析器来说是非常有利的一面。

- 非规范语句存在比率较大。在上述我们统计的语料中，如果将独词句排除在外，仅算重复、次序颠倒、冗余和省略4种情况，累计出现的几率就高达42.1%，这些非规范语句将是口语理解机制的主要障碍。

- 非文字信息的丢失导致句型歧义。由于口语转化成文字以后，许多非文字信息（语气、停顿、声调等）丧失，使本来没有歧义的语句具有了歧义。因此，口语识别中的韵律特征识别和声调识别显得非常重要，子句边界的识别和语句切分也变得十分重要。

尽管我们处理的语料限定在特定领域，而且也是非常有限的，但是，这些语料基本上反映了人们日常口语对话中用词、造句的基本规律。因此，本文统计的结果对于不同领域的口语对话翻译系统和人机对话系统研究都是具有较大的参考价值。

参考文献

- [1] Rebecca J, Passonneau, Diane J. Litman., "Discourse Segmentation by Human and Automated Means", Computational Linguistics., Vol. 23, No. 1, 1997, Pages 103~139.
- [2] Marilyn A, Walker, Johanna D, Moore, "Empirical Studies in Discourse", Computational Linguistics. Vol. 23, No. 1, 1997. Pages 1~12.
- [3] Alexandra Georgakopoulou, Dionysis Goutsos, "Discourse Analysis", Edinburgh University Press, 1997.
- [4] 陈建民, "汉语口语", 北京出版社, 1984.
- [5] 刘源, 谭强, 沈旭昆, "信息处理用现代汉语分词规范及自动化分词方法", 清华大学出版社, 1994.
- [6] 丁信善, "语料库语言学的发展及研究现状", 当代语言学, No. 1, 1998. Pages 4~12.