基于口语度的口语词语自动提取研究

侯敏 张玉强 何伟 邹煜 滕永林
中国传媒大学 国家语言资源监测与研究中心 有声媒体语言分中心 北京 100024
E-mail: houminxx@263.net

摘要：口语词语自动提取的最大障碍在于口语语料的难以获取和口语词语界定的模糊性。本文充分利用广播电视语料兼具书面语体和口语语体的特点，提出了口语度计算模型，该模型以 Logistic 回归模型为基础，以词语空间分布通用率为协变量，通过衡量词语在书面语体语料和口语体语料中的空间分布差异，能够有效地度量该词语的口语度，从而实现口语词语的自动提取。在约 1100 万字语料上的实验结果表明，口语和书面语共现词语中提取口语词语准确率为 85%，口语独现词语中提取口语词语准确率为 76.5%，平均正确率达到 79.3%。

关键词：口语度 口语词语 自动提取 Logistic 回归模型

Automatic Extraction of Spoken Words by the Spoken Language Measurement

Hou Min, Zhang Yuqiang, He Wei, Zou Yu, Teng Yonglin
Broadcasting Media Language Research Center, Communication University of China, Beijing 100024
E-mail: houminxx@263.net

Abstract: Lack of spoken corpus and the ambiguity definition of spoken words is the biggest obstacle to automatic extraction of spoken word. This paper used the broadcasting corpus which comprised both written and spoken language, and proposed the spoken measurement calculation model. The model is based on Logistic Regression Model with the words generalization as covariates, which could measure the differences of spatial distribution between words in the written corpus and that in spoken corpus, thus the probability of spoken words can be effectively measured and the spoken words can be extraction automatically. The results of experiments on about 11 million words show the precision of extraction is 85% for the words occurred in both spoken and written language and 76.5% for the words occurred only in spoken language, the total precision is 79.3%.

Keywords: spoken language measurement, spoken words, Logistic Regression Model

1 引言

口语词语是形成口语语体的基本要素之一，也是语言研究的重要内容。口语词语的总结、归纳对于口语理解、语音识别等语言工程开发具有重要意义。但是，相对于书面语研究来说，汉语口语研究一直是个薄弱环节，其原因有口语语料的获得和保存十分困难，口语词语概念的模糊性，口语词语没有明显的形式特征等。

传统口语研究大多集中在口语语法上，即使是对口语词语进行研究，也多是以作家作品等书面文本为基础，其语言素材规模不大，研究范围相对较窄。随着语料库研究方法的引入，汉语口语研究得到了丰富和发展。文献[1]通过“当代北京口语语料库”对北京话高频词的使用状况进行了分析。文献[2]等利用语料库，对汉语口语的关系从句的分布进行了统计分析研究。文献[3]从电话交谈中选取大量语料，考察了会话中“对吧”的分布位置，总结了“对吧”在会话中的语用功能。但在口语词语自动提取方面尚未发现相关报道和文献。

本文以有声传媒语言语料库[4]为基础，充分利用广播电视语料语体混杂的特点，分别抽取文稿播报语料和谈话语料，构建了共时的书面语语料库和口语语料库，通过对比词语在不同语体语料中的空间分布差异，提出口语度指标加以度量，从而达到自动提取目的。为此，本文提出词语空间分布通用率模型来表示词语的空间分布，并引入 Logistic 回归模型，以词语在不同语体空间的通用率为协变量，得到口语度计算值，以此为依据进行提取实验，并对实验结果进行了人工评价的语感实验与分析。以下第二节将介绍词语空间分布通用率模型，第三节介绍口语度计算，第四节为实验及分析。

2 词语空间分布通用率模型的建立

如何选择恰当的可计算因子以构建词语空间分布通用率模型，需要结合语言事实的考察，下表是选择的部分词语在语言事实中的分布情况。

表 1 词语在语料中的分布示例

词语名称	书面语料出现频次	书面语料文本数	书面语料栏目数	口语语料出现频次	口语语料文本数	口语语料栏目数
出席	1744	609	5	6	6	4
发布	1512	511	5	110	66	15
机制	961	451	5	144	65	9
值班	73	56	5	63	37	10
清晰	77	70	4	84	63	16
接触	115	103	5	382	219	17
反正	4	4	3	606	348	17
当中	152	124	5	2472	468	17
刚才	31	30	4	2690	529	17

表 1 中“出席、发布、机制”三个词语书面语色彩比较浓，在书面语中的分布频次都非常高，文本分布比例为 63%，53%，47%，在所选择的 5 个栏目中均出现。而这三个词语在口语语料中的分布就相对比较少，文本分布比例分别为 0.5%、5.8%、5.7%，栏目分布比例分别为 24%、88%、53%。因此可以得出，这三个词语绝大多数的时候是用在书面语体中，这与人们的语感是符合的。

“值班、清晰、接触”三个词语是通用体词语，从表中可以看到，这三个词语在口语和书面语体中的总体分布比较接近。

而第三组词语“反正、当中、刚才”具有比较强的口语色彩，从表中看到，它们在在口语语体语料中，出现的频次高、分布广、而且出现在所有的栏目中，而在书面语体语料中，与口语体语料中分布相比，这三个词语，频次低，分布小，所以它们应该比较接近口语。

因此，传媒语料库中，词语的分布可以通过词语的使用频次、文本数、栏目数三个方面来描述[5]。

通过对语料的分析，我们认为，应该用词语在语料中出现的频次、文本数、栏目数三个变量参数来描述词语的空间分布特征；基于此，本研究提出了传媒语言的归一化通用率模型如下：

$$\Phi_t = U_t = \frac{C_w \times tf_w \times df_w}{\sum_{w \in V}(C_w \times tf_w \times df_w)} \quad (1)$$

其中，U_t 为归一化通用率，C_w 为词 w 的频次，df_w 为该词的文档分布率，tf_w 为该词的栏目分布率，分母为归一化项，V 表示所有词种。

tf_w，df_w 计算公式如下：

$$tf_w = t_w / T \quad (2) \qquad df_w = d_w / D \quad (3)$$

其中，d_w 为含有词语 w 的文本数，D 为文本总数；tf_w 为含有词 w 的栏目数，T 为栏目总数。

3 口语度计算

对语言现象进行计算，属于定量分析的范畴。口语词语是一个模糊概念。我们试图通过计算口语度，来判别一个词语是否属于口语词语或者该词语在多大程度上属于口语词语，即它属于口语词语的概率是多少。我们模型中的因变量是一个典型的二元分类变量。“在分析分类变量时，通常采用的一种统计方法是对数线性模型。”“当对数线性模型中的二分类变量被当作因变量并定义为一系列自变量的函数时，对数线性模型就变成了 logistic 回归模型” [6]。在本文中，确定一个词语是否属于口语词语，依据的就是词语的空间分布特征，也就是说，因变量（其值为 1 或 0，1 表示是口语词语，0 表示是非口语词语）的值是由自变量（也称为协变量，本文指词语分别在口语和书面语料中的空间分布通用率）决定的。所以 logistic 回归模型完全适用于本研究。由于 logistic 回归模型在不同情况下有不同的数学符号表示形式，为方便起见，本研究统一使用 logistic 回归模型表示形式为：

$$P(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)} \quad (4)$$

其中，Y 为因变量，取值 1 和 0，1 表示词语属于口语词语，0 表示词语是非口语词语；x_1，x_2，……，x_q 表示自变量；β_0，β_1，β_2，……β_q 是模型待确定的参数。P 表示 x 属于 1 的概率，也就是我们所要计算的口语度。

4 口语词语提取实验

本文将口语词语划分为两类，即“共现口语词语”与“独现口语词语”。“共现口语词语”指的是在口语语料和书面语料中都出现过的口语词语，“独现口语词语”，指的是只在口语语料中出现的口语词语。针对这两类口语词语分别建立模型并提取，实验过程包括语料准备，种子集筛选与模型建立，确立阈值与词语提取等步骤。以下分语料准备、共现口语词语提取、独现口语词语提取、人工评价的语感实验与分析四小节分别叙述。

4.1 语料准备

广播电视语言既不是典型的日常生活口语，也不是典型的书面语，它在总体上是介于口语和书面语二者之间的一种比较规范的语言[7]。因此，就需要我们从传媒语言语料库中选择尽可能接近口语和尽可能接近书面语的语料分别作为口语和书面语的代表。

本文中选择了《实话实说》、《对话》等 17 个栏目共 560 多万字作为口语语料，选择了《新闻联播》、《北京新闻》等 5 个新闻栏目 600 万字作为书面语语料。为保证实验效果，在实验前对语料进行了简单的预处理，比如把《新闻联播》等栏目中的“同期声”去除掉等，预处理后的二者规模大体相当。

4.2 共现口语词语提取

经统计，数据库中共有共现词语 27013 条记录，共现口语词语将从其中提取。

第一步：种子测试集合的选取

挑选种子测试集时坚持：要使测试集尽可能有代表性，比如既要同时考虑到高、中、低频次词语，又要选择各频次词语中空间分布特征差异较大的词语。我们以口语语料中词语的频次为依据，把第一部分语料分为三个频次段，每段都选取一部分比较典型的词语作为种子进行测试。经过认真挑选，挑选出口语词语种子 217 个，书面语词语种子 232 个。

第二步：参数拟合及结果分析

把种子测试集中词语的通用率值作为协变量代入 Logistic 二元回归模型，再次进行参数值的拟合。参数拟合结果的显著度等于 0，常数量的参数值也仅为 0.007，因此模型拟合成功。

在模型参数的拟合过程中，以 su 表示书面语体中词语空间分布通用率，以 ku 表示口语体中词语空间分布通用率。我们发现协变量 su、ku 在计算过程中其值非常小，因此在使用时对其进行了简化，去掉了分母这个固定值，简化得到：

$$su = samount \times stime / ST \times sdir / SD$$

$$ku = kamount \times ktime / KT \times kdir / KD$$

其中，Samount、stime、sdir、kamount、ktime、kdir 分别表示词语在书面语料中出现频次、文本数、栏目数和在口语语料中出现频次、文本数、栏目数。ST SD KT KD 分别表示书面语语料中的文本总数、栏目总数和口语语料中的文本总数和栏目总数。

第三步：把迭代历史中最后一次迭代的参数值代入 logistic 回归模型公式，得到共现词语口语度计算模型如下：

$$P(Y=1 \mid X) = \frac{\exp(0.477 - 0.107su + 0.014ku)}{1 + \exp(0.477 - 0.107su + 0.014ku)} \qquad (5)$$

第四步：依据公式 5，分别计算每个词语为口语词语的概率。

第五步： 确定阈值，提取口语词语。

我们对 27013 条纪录在不同阈值之间的分布进行了统计，结果如表 2。我们知道，在所有的词语中，具有口语色彩的词语相对是比较少的，更多的是通用体词语。在表 2 中，阈值 P 在 0.6 到 0.7 之间的词语特别多，占据了整个数据 80%以上，我们对这部分词语的分布再做统计，最后，确定口语色彩比较重的词语的口语度 P 为大于等于 0.62。根据口语度这个阈值，从共现词语中提取了 1024 条口语词候选。

4.3 独现口语词语提取

经统计，数据库中共有独现词语 12975 条记录，独现口语词语将从其中提取。

从理论上说，这些独现词语口语度都应该为 1。但分析发现，这些只出现在口语语料中的“独现词语”频次比较高的主要是大量的人名和少量的口语词，而且空间分布特征参数“栏目数”差异较大。比如某些行业用语只出现在一个栏目的一篇文档当中，而且出现的频次也非常低；某些主持人的名字，虽然也只出现在 1 到 2 个栏目，但文本数和频次都非常高。下面对这部分词语应用 Logistic 回归模型进行分析。

首先，选择种子测试集，并把其中的口语词语的自变量（y）值设为 1，人名及非结构词语等自变量设为 0。挑选口语种子 145 个，非口语种子 148 个。以频次（kamount）、文本数(ktime)、栏目数(kdir)为协变量，y 为自变量，进行 Logistic 回归模型参数拟合。拟合结果的显著度接近于 0。

于是，我们得到独现词语口语度计算模型公式：

$$P(Y=1\mid X)=\frac{\exp(0.656-0.110\,kamount+0.110\,ktime+0.661\,kdir)}{1+\exp(-0.656-0.110\,kamount+0.110\,ktime+0.661\,kdir)} \quad (6)$$

按照概率值对 12975 个词语进行统计，得到分布结果如表 3。对于口语独现词语，我们根据表 3，并对实际的语料进行了考察，确定了阈值为 P1≥0.8。根据这个阈值，从独现词语中提取了 2105 条口语词候选。

表 2 共现词语口语度 P 值不同范围词语分布表

口语度 P	词语个数	所占比例
P=1	28	0.1%
0.9≤ P <1	107	0.4%
0.8≤ P <0.9	54	0.2%
0.7≤ P <0.8	98	0.4%
0.6≤ P <0.7	21835	80.8%
0.5≤ P <0.6	2525	9.3%
0.4 ≤ P <0.5	610	2.3%
0.3≤ P <0.4	402	1.5%
0.2≤ P <0.3	236	0.9%
0.1≤ P <0.2	255	0.9%
0 ≤ P <0.1	845	3.1%
P=0	18	0.1%
总计	27013	100%

表 3 口语独现词语以 P1 为区间的分布

概率 P1	词语个数	所占比例
P1=1	0	0
0.9≤ P1 <1	2105	16.22%
0.8≤ P1 <0.9	2540	19.58%
0.7≤ P1 <0.8	7833	60.37%
0.6≤ P1 <0.7	129	0.99%
0.5≤ P1 <0.6	62	0.48%
0.4 ≤ P1 <0.5	62	0.48%
0.3≤ P1 <0.4	29	0.22%
0.2≤ P1 <0.3	33	0.25%
0.1≤ P1 <0.2	75	0.58%
0 ≤ P1 <0.1	144	1.11%
总计	12975	100%

4.4 人工评价的语感实验与分析

口语词语本身就是一个模糊的概念，没有精确的界定，所以无法进行严格的准确率和召回率验证。为考察模型的可信度和准确率，我们做了一个人工判别的实验。从提取结果中按10%的比例随机选择词语，共抽取词语300个，然后请15名语言学专业和中文专业人员分别为每个词语评分。评分标准是：典型的口语词语10分，典型的书面词语0分，典型的通用体词语5分。请实验人员根据自己的语感打分。为了保证实验结果的客观性，我们在做实验时只给出了词语和评分标准，没有给出计算的口语度值以及各个参数。表4是统计结果。

表4 人工评价的语感实验结果

类 型	评分均值≥5	比例（%）	评分均值< 5	比例（%）
共现口语词语（100）	85	85	15	15
独现口语词语（200）	153	76.5	47	23.5
总计（300）	238	79.3	62	20.7

对口语度各个值区间与相应人工评分均值的主客观关联分析表明，口语度降低，人工评分值缩小，口语度与人工判别评分在总体趋向上一致。

独现词语中包含的口语词语多，正确率却比共现词语的提取结果低，主要原因在独现词语中，低频词过多，对于同是低频词语的口语词语或通用体词语，其空间分布特征很难区分，导致准确率有所下降。

值得注意的是，在这3129条口语词候选中，有相当一部分是《现代汉语词典》以及其他词典中没有给出的活生生的口语词语。如“说白了、不好说、说真的、一般来讲、这么着（zhèmezhāo）、不是说、弄不好、大体上、硬着头皮、不相干、这一来”等等。这些词语在人们口头使用频率很高，但一般词典都没有收条，给语言教学，尤其是对外汉语教学带来很大的困难。能把它们提取出来，就为下一步的口语词典编纂奠定了基础。

5 结语

本文在充分借鉴术语、流行语等词汇自动提取及语言计量研究成果的基础上，提出了口语度计算模型，利用广播电视语料特点，在大规模语料库的基础上，进行了口语词语自动提取实验，并对提取结果进行了人工判别的语感实验。实验结果表明，提取结果与人们的语感基本趋向一致，证明了口语度计算模型用于口语词自动提取的可行性和客观性，为口语词语的进一步大规模提取奠定了基础。

参 考 文 献

[1] 高海洋，北京话高频词使用状况分析，中国社会语言学，2003(1)：26-33.

[2] 方梅，口语语法研究的现状与前瞻，21世纪的中国语言学（一），北京：商务印书馆，2004.

[3] 梁丹丹，会话中“对吧”的语用功能，修辞学习，2006(1).

[4] 侯敏等，传媒有声语言语料加工体系研究，中国传媒大学211工程项目研究报告，2006.3.

[5] 何伟，侯敏等，流行语时空监测模型的研究，内容计算的研究与应用前沿，清华大学出版社，2007.

[6] 王济川，郭志刚，Logistic回归模型——方法与应用，高等教育出版社，2001.

[7] 赵雪，关于广播电视语体的思考，现代传播，2000，104(3)：98~99.