基于词语聚类的汉语口语教材自动推送素材研究[1]

杨冰冰
北京语言大学汉语学院
942249583@qq.com

赵慧周
信息科学学院
zhaohuizhou@blcu.edu.cn

王治敏*
汉语国际教育研究院
wangzm000@qq.com

摘要

新冠肺炎的蔓延使得线上移动教学成为教育发展的必然趋势，本文以适合汉语教材自动推送的口语素材为研究对象，基于10341条生活类口语语料，对词汇的整体特点进行计量分析，在此基础上根据腾讯AL LAB 公开的中文词向量数据，使用Kmeans算法对口语词汇进行词语聚类，参考词语聚类结果及对口语语料话题和场景的考察，构建了一个包含15个一级话题、102个二级话题及81个交际场景的汉语口语话题-场景素材库。同时对各级话题常用词进行了总结。本文可为教材自动定制的素材库提供资源支持。

关键词： 汉语资源建设；词语聚类；教材自动推送；汉语口语

Study on Automatic Push Material of Oral Chinese Textbook Based on Word Clustering

Yang Bingbing
Beijing Language and Culture University Chinese Language Institute
942249583@qq.com

Zhao Huizhou
Beijing Language and Culture University School of Information Science
zhaohuizhou@blcu.edu.cn

Wang Zhimin
Beijing Language and Culture University Chinese International Education Research Institute
wangzm000@qq.com

Abstract

The spread of Coronavirus has made online teaching an inevitable trend. This paper takes oral materials suitable for automatic push of Chinese textbooks as the study object. Based on 10341 daily oral sentence, the overall characteristics of the vocabulary are quantitatively analyzed. On this basis, according to the Chinese word vector data published by Tencent AL LAB, the Kmeans algorithm is used to analyze Spoken language vocabulary is clustered, referring to the results of word clustering and the investigation of the topics and scenes of the spoken language corpus, a Chinese spoken topic-scene material library containing 15 first-level topics, 102 second-level topics and 81 communication scenarios is constructed. At the same time, it summarizes the commonly used words on topics at all levels. This paper can provide resource support for the material library automatically customized for teaching materials.

Keywords: Chinese resource construction , word clustering , textbook automatic push , spoken Chinese

[1]本研究得到国家社科基金重大项目“基于‘互联网+’的国际汉语教学资源与智慧教育平台研究”（18ZDA295），中央高校基本科研业务费（18YBT03；19PT03）的资助。

第十九届中国计算语言学大会论文集，第624页-第633页，海南，中国，2020年10月30日至11月1日。

1 引言

新冠肺炎在全世界的蔓延使得在线学习成为一种趋势。在线学习资源如何满足学习者的学习需求，如何为学习者提供丰富的、可供选择的个性化学习素材，如何联通汉语学习者与教师这两部分群体，是目前特殊大环境下亟待解决的问题，也将是语言教育创新发展，适应互联网时代发展面临的问题。

本文面向在线学习中的汉语口语教材自动定制，对其所需要的学习素材进行研究。教材自动定制需要联通词汇、场景、话题等模块，构建相应的推理模式。关于词汇资源的研究，目前很多词汇知识库的研究主要是基于规则和基于统计的，也有学者利用文本分类中特征提取的方法在大规模分类语料中自动获取领域词语（刘华，2007）。基于词语聚类，一些学者对不同话题领域的词表进行了研究（吕荣兰2011；喻雪玲2013；陈钰铭2015）。关于对外汉语中的话题分类，《国际汉语教学通用教程大纲》（2008）列出了包括22个话题以及其下的小话题与常用表达，一些学者构建了汉语话题库（吕荣兰2011；苏新春2011；方沁2014）。目前对汉语交际场景的研究有杨寄洲（2000）在《对外汉语教学初级阶段教学大纲》中列举的36个交际场景，其后刘华（2014）根据影视片段，总结了63个交际场景。现有研究对我们的研究有一定的参考作用，但是目前对生活类话题下的二级话题的研究还不够深入，且缺少对交际场景和话题关联的相关研究，对汉语口语的特点也缺少分析。

因此本文以适用于汉语教材自动推送的教材素材为研究对象，通过对口语词汇特点的分析以及词语聚类，总结口语话题及场景，构建话题-场景对应的口语素材库，为在线口语教材自动推送提供资源支持。

2 汉语口语词汇特点分析

本文以汉语生活类口语为切入点，收集了10341句生活类语料。语料来源于《英语会话8000句》中生活类话题下的句子以及真实场景的对话录音。《英语会话8000句》作为英语口语的代表教材，经过长期的使用可以证明其话题覆盖面较宽，句子实用性较强，且与汉语口语教材有一定的区分，经过筛选与改写后可以作为汉语口语教材的原始语料。关于口语录音，本文对生活中常见话题、场景进行了真实场景的录音，通过机器转写与人工校对形成本文的另一部分基础语料。笔者所处语言环境为北京的高校，录音场所为北京内的交际场所。通过对全部语料进行机器分词和人工校对[2]，去除数字、英文字母后共提取出词汇6803个。

2.1 词类分布特点分析

在对全部词汇进行分词与词性标注的基础上，我们发现在口语会话中各类词汇分布如图1：

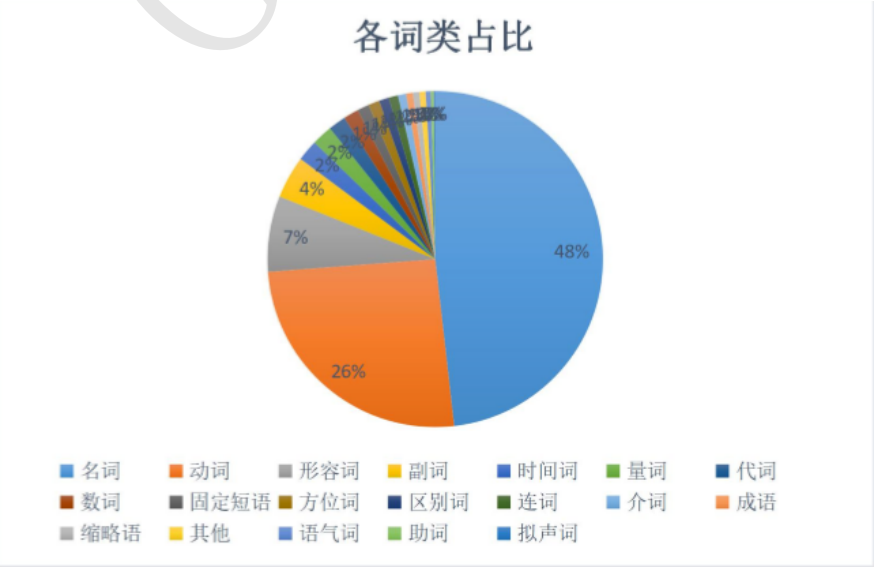

图 1 各词类分布图

[2]分词标准参考《现代汉语语法信息词典》

图1表示在所有口语语料中各词类丰富度情况，总体来看，在口语会话中名词种类最多，数量为3279个，占全部词汇的48%。其次为动词，数量为1741个，占比26%，再者为形容词，数量为496个，占比7%，副词数量为278个，占比4%。这四种词类包含的不同词语种类最多，共占全部词汇的85%，与汉语词类构成相符。时间词、量词、代词数量丰富，数量分别为139个、135个、115个，体现出口语对话中用词的特点。在口语对话中存在一部分固定短语，如“不好意思、没关系、可不”等，数量为78个。另外，在口语语料中也有一部分成语，但是数量不多，为44个。另外有一些无法判定词类的词语我们将其分为“其他”类。从各词类丰富度来看，名词、动词、形容词依旧是口语学习的重点，语气词、助词等属于封闭类词类因此数量不多，但是重点词语较多，如语气词中的“吧、呢”；助词中的“的、了”等。

词频反映一定范围内词语的常用度，因此下文将对口语对话中各词类的词频进行统计分析。

图 2　各词类词频图

由图2可知，各词类词频与各词类词语数量不一致，在口语语料中，实词的出现频率高于虚词。在所有词类中动词的词频最高，总词频为19718次，由此可知在口语对话中动词的常用度在所有词类中占据优势。其次为代词，代词虽数量不多，但是出现的频次很高，口语对话大多是两人之间的对话，因此进行自我阐述与询问对方使用的人称代词出现的频率很高。词频第三为名词，总频次为10909次，名词数量为3191个，这反映出大部分名词出现频次较低，常用度较高的名词不多。词频第四为语气词，频次为6132次，语气词是虚词中频次最高的词类。语气词虽数量不多，但是出现频率很高，语气词的高频使用反映了汉语口语的特点。其次为助词，频次为4598次，使用频率较高。另外在口语语料中，成语、拟声词等词频较低，常用度较低。

2.2　高频词词汇特点分析

口语中的高频词可以体现口语词汇特点，在口语中词频最高的词为：“我、的、你、了、是、好、吗、不、吧、您”等。在口语中代词“我”的频次最高，这与《中国语言生活研究报告》中“的”词频第一不一致，因为在口语对话中多为表达自己观点的句子，因此代词“我”的词频比书面语等语体中高。另外其他人称代词“你”、“您”、“我们”的词频也很高。

从音节数来看，在词频为前50的词中，大多为单音节词，数量为41个，双音节词数量较少，仅为9个。单音节词的数量远多于双音节词及多音节词，且频率越高，单音节词优势越强。从难易程度上看，高频口语词汇中大部分属于汉语低水平词汇，通俗性较强。

为了对高频词的词类分布进行下一步的研究，我们对词频前50各词类数量进行了统计，如图3所示。

由图3可知在高频词中动词数量最多，共15个，其次为代词，数量为10个，副词和语气词的数量也较多。助词、介词、数词中各有2个高频词，形容词、方位词、名词、量词、时间词中各有一个高频词，高频词数量较少。其他没有列出的词类没有频次在前50的高频词。

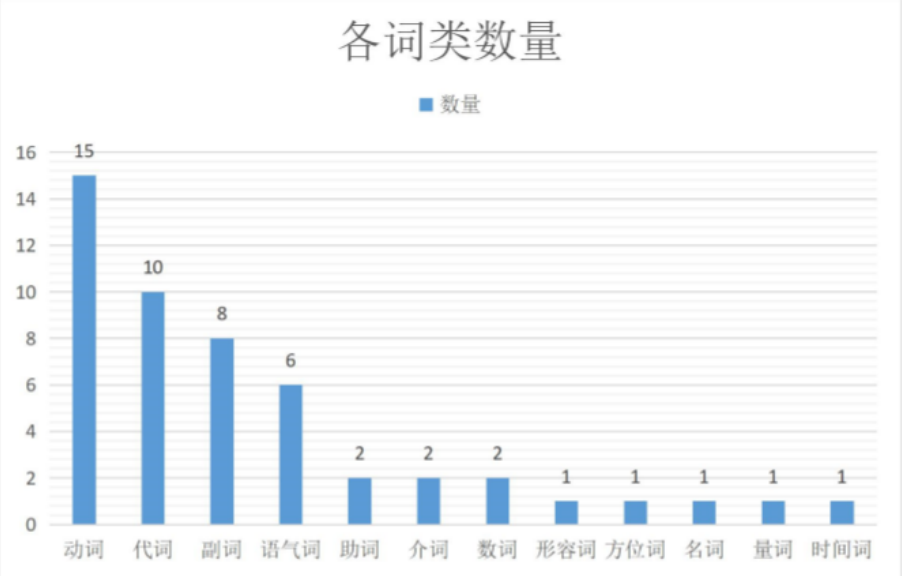

图 3　频率前50词类统计图

3　基于词向量的词语聚类

词语聚类可以把语义特征相似的词语聚在一起，自动定制的汉语教材需要提供各话题、场景的常用词及常用句，词语聚类结果可为话题和场景下常用词的汇总提供参考。本文对全部10341条语料进行话题的场景的标注，在词语聚类和标注的基础上归纳生活类话题和场景。

3.1　词向量模型

文本的向量化即把文本转化为n维的向量。本研究使用腾讯AL LAB 公开的中文词向量数据，包含800万词汇，每个词对应一个200维的词向量，该词向量数据包含很多现有公开的词向量数据所欠缺的短语，所计算的语义相似度较高，且采用了Directional Skip-Gram(DSG)算法作为词向量的训练算法，DSG算法基于广泛采用的词向量训练算法Skip-Gram(SG),在文本窗口中词对共现关系的基础上，额外考虑了词对的相对位置，所生成的词向量能够更好地表达词之间的语义关系。

3.2　Kmeans聚类算法

Kmeans 算法是最经典的基于划分的聚类方法，首先确定想要经过聚类得到若干集合的k值，我们将k值设置为500、800和1000。从全部的词汇向量中随机选择k个数据点作为质心。计算所有向量数据与聚类中心的相似度，距离离质心越近，相似度越高，计算公式如下：

$$|AB| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

计算完成后将数据分配给与其最相似的聚类中心代表的类，并计算下一轮聚类的中心。如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值，我们可以认为聚类已达到期望的效果，算法终止。如果新质心和原质心距离变化很大，需要继续迭代。聚类为500类、800类及1000类的实验结果如下，以生病就医类词语为例。

由表1可知，聚类总数越多，同一话题内词汇的聚类数也越多，分类越细致，每类之间的区分度越弱。在聚类数为500类的词语聚类中，对药物名称、药物效果及大夫等的划分包含在了一个类里，而在聚类数为800类的聚类中，药物名称单独为一类，在聚类数为1000的聚类中，药物名称根据种类又分为了两个类。因此对于口语话题来说，对于大话题的划分参考聚类数为500的词语聚类，对于下级话题和场景的划分参考聚类数为1000和800的词语聚类结果。

聚类数	平均词语数量	词语
500类（10类）	10.80	医院、出院、报销、住院、护士、看病、家属、病房、医保、转院、陪床、动手术、查房、住院处、住院费、主刀、药费
		医生、手术、病、治疗、病人、化疗、疗程、患者、微创、会诊、切除、预后、术后、病情、病理、不治之症、康复、确诊、疗效、治愈率、癌症、就诊、矫正、输血
		感冒、发烧、生病、咳嗽、流感、重感冒、鼻塞、麻疹、传染、退烧、上火、发高烧、感染、着凉
		血、伤口、伤、受伤、绷带、清创、切口、外伤、愈合、流血
		药、大夫、药方、治、中药、药店、处方、安眠药、药物、口服、盘尼西林、见效、药效、药膏、阿司匹林、送服、药品、体温计、方子、药水
		吃药、忌口、牙疼、打针、戒烟、服药
		药房、配药、化验室
		挂号、挂号费、挂号单、专家号、血压、血糖、心率、心电图
		眼科、诊室、内科、中医科、科室、牙科、北医三院
		账单、收据、收付款、票据、预约单、化验单
800类（15类）	7.87	医院、出院、住院、护士、救护车、眼科、会诊、转院、牙科、内科、化验室、科室
		病房、诊室、查房、住院处、中医科
		医生、大夫、病人、家属、患者、肺、病理
		病、感冒、发烧、照顾、生病、吃药、病倒、体质、打针、静养、重感冒、陪床、卧床、养病、发高烧
		嗓子、咳嗽、闷、喘
		流感、传染、感染
		治疗、化疗、治、病情、康复、确诊、退烧、靶向、疗效、治愈率、矫正
		感康、止痛片、布洛芬、感冒药、头孢
		伤、受伤、后遗症、伤筋动骨
		疼、肿、痛、疼痛、痒、剧痛、蛰、血压、血糖、心率、心电图
		手术、微创、切除、开刀、拆线、术后、动手术、输血、主刀
		药、中药、西药、中成药、药物、药方、含片
		服用、疗程、口服、阿司匹林、送服、服药
		看病、挂号、挂号费、预约单、挂号单、化验单、转诊单、就诊、专家号
		报销、医保、住院费、药费、医保卡
1000类（21类）	5.62	医生、大夫、医院、病人、护士、看病、家属、会诊、诊室、动手术、查房、就诊、主刀
		眼科、牙科、病理、内科、中医科、科室
		感冒、发烧、生病、病倒、重感冒、发高烧
		流感、传染、麻疹、感染
		嗓子、咳嗽、鼻涕、闷、咳、喘、憋、呛
		疼、酸痛、肿、痛、疼痛、剧痛
		伤、受伤、事故、崴脚
		伤口、绷带、清创、外伤、愈合、流血

第十九届中国计算语言学大会论文集，第624页-第633页，海南，中国，2020年10月30日至11月1日。

聚类数	平均词语数量	词语
		治疗、护理、化疗、患者、预后、术后、病情、康复、确诊、治愈率、矫正
		吃药、打针、拆线、麻药
		药物、副作用、药品、疗效、激素、麻醉剂
		药、中药、配药、西药、中成药、安眠药、止痛片、药方、含片、退烧、止咳
		布洛芬、感冒药
		感康、头孢
		药房、药店、器械公司
		手术、微创、切除、切口、阑尾、活检
		手术、微创、切除、切口、阑尾、活检
		静养、养病
		出院、住院、病房、转院陪床
		挂号、挂号费、预约单、挂号单、住院处
		报销、医保、住院费、药费

表 1　500类、800类、1000类词语聚类情况

4　话题—场景库的构建

4.1　话题的确定

基于词语聚类结果，在对12本常用口语教材、《国际汉语教学通用教程大纲》（2008）、吕荣兰（2011）、方沁（2014）等话题库以及英语教材的话题进行总结的基础上，我们总结出生活类一级话题15个，分别为："个人信息、日常交际、居家生活、运动健身、生病就医、交通出行、购物、饮食就餐、天气、日期时间、资金管理、生活服务、休闲娱乐、意外与事故和住宿"。每个一级话题下又有若干个二级话题，根据语料的场景标注，每个二级话题都有其对应的场景。以"生病就医"话题为例：

一级话题	二级话题	场景
生病就医（I）	不适	家（JA）、宿舍（SS）、户外（HW）、路上（LS）、办公室（BG）
	预约	电话（TP）、预约平台（YP）、医院（HP）
	挂号	医院（HP）、预约平台（YP）
	就诊	医院（HP）、诊室（ZS）、诊所（ZS）、病房（BF）
	费用	医院（HP）
	买药	药房（YF）
	手术	手术室外（OR）、诊室（ZS）、病房（BF）
	住院	病房（BF）、医院（HP）、住院部（ZY）
	探病	病房（BF）、住院部（ZY）、家（JA）

表 2　"生病就医"二级话题、场景

参考词语聚类结果，我们总结出上表所示在一级话题"生病"中有二级话题9个，二级话题排列顺序基本遵循去医院看病流程。总体来看，关于"生病就医"话题的对话场景较为固定，几乎所有的二级话题中涉及到的场景都有"医院"。

通过对全部一级话题对话及词语聚类的考察，本文总结出15个一级话题下共102个二级话题，每个话题下有其对应的交际场景；根据对话内容，总结出每个话题和场景下的口语常用句。

4.2　各级话题词汇特点及常用词分析

不同话题下词汇具有不同的特点，教材自动推送资源需要提供各个话题、场景下常用度高、代表性强的词汇。为提高话题词汇相关性，突出各话题词汇特点，本文对各级话题词汇特

点及常用词的研究去除总词频中频次最高的前15个词，分别是：我、的、你、了（语气词）、是、好、吗、不、吧、您、这、有、就、我们、在，且去除一些无意义的虚词，分别是助词、介词、连词和语气词。我们对每个一级话题下的词汇特点进行了分析（以“生病就医”为例）：

词类	词频	词语及频次
动词	645	要（69）可以（56）去（54）做（48）吃（42）能（36）看（32）来（32）需要（30）想（29）休息（24）说（24）开（23）拿（22）得（20）手术（18）感冒（16）治疗（13）出院（11）预约（11）挂号（11）发烧（9）服用（8）住院（7）
名词	185	药（42）医生（34）时候（23）时间（23）大夫（18）医院（16）药房（11）体温（10）病人（8）
形容词	63	疼（27）多（16）舒服（12）严重（8）

表 3　“生病就医”话题高频词统计

通过统计发现在“生病就医”话题中高频动词数量很多，其中“休息、开、手术、感冒”等话题相关性较强，名词中“药、医生、大夫”话题相关性较强，形容词中“疼、舒服、严重”具有话题代表性。可作为“生病就医”一级话题下的常用词参考。

二级话题	场景	常用词
不适	家、宿舍、户外、路上、办公室	怎么了、哪里、疼、有点、药、医院、看病、感冒、难受、发烧、头疼、脸色、生病、医生、着凉、肚子、受伤、舒服、咳嗽、胃口
预约	电话、预约平台、医院	预约、时间、有空、合适、上班、满、网上、公众号
挂号	医院、预约平台	挂号、专家号、普通号、预约、网上、科、交费、挂号费、排队、内科、外科、口腔科
就诊	医院、诊室、诊所、病房	怎么了、哪里、问题、疼、药、医生、大夫、治疗、做、检查、严重、开药、休息、服用、情况、打针、输液、症状、伤口、医院、拍片、化验、量、注意、饮食
费用	医院、交费处	报销、医保卡、刷卡、单子、交费、交费处、排队、现金、自助
买药	药房	中药、药方、消炎、退烧药、过敏、配药、次、片、粒、过敏、症状、头孢、感冒、感冒药、止疼片
手术	手术室外、诊室、病房	手术、签、术前、家属、安排、术后、体质、风险、肿块、同意书、营养、恢复、开刀、部位、麻醉、麻药、输血、微创
住院	病房、医院、住院部	住院、出院、注意、休息、办、手续、陪床、检查
探病	病房、住院部、家	休息、怎么样、放心、情况、担心、营养、出院、照顾、保重、早日康复

表 4　“生病就医”二级话题常用词

在对每个一级话题进行词频统计的基础上，我们提取出二级话题中的话题高频词，并参考高频词的词语聚类结果，总结各二级话题下的常用词：

4.3　生活类场景统计分析

通过对语料场景的标注以及对话题的考察，我们统计出81个生活类场景，相比较《对外汉语教学初级阶段情景大纲》（杨寄洲，2000），涵盖的生活范围进一步增加，场景是口语交际发生的场所，培养学习者根据场景进行交际是培养口语交际能力的关键。口语素材库需要尽可能多的交际场景来满足学习者的交际需求。具体生活类场景如表5：

[3]括号内数字表示该场景在多少个一级话题中出现。

场景类型	场景
居住场所	家（12）[3]宿舍（10）小区（3）厨房（1）宾馆（1）
交通工具	出租车（2）地铁（1）地铁站（1）公交车（1）公交站（1）火车（1）火车站（1）飞机（1）机场（1）
就餐场所	饭店（3）咖啡厅（3）食堂（2）酒吧（2）快餐店（1）西餐厅（1）火锅店（1）奶茶店（1）小吃街（1）甜品店（1）
运动场所	操场（3）体育馆（2）篮球场（2）足球场（2）健身房（1）瑜伽馆（1）游泳馆（1）
购物场所	商场（3）市场（2）超市（1）商店（1）服装店（1）家具店（1）书店（1）
就医场所	医院（3）诊所（1）医院诊室（1）病房（1）住院部（1）手术室外（1）药房（1）
休闲娱乐场所	电影院（1）剧院（1）KTV（1）音乐厅（1）游乐场（1）景区（1）赛车场（1）滑雪场（1）美容院（1）美甲店（1）游戏厅（1）售票处（1）
学习办公场所	校园（7）办公室（5）公司（4）教室（3）图书馆（2）
服务场所	维修店（1）洗衣店（1）理发店（1）快递点（1）打印店（1）
公共场所	路上（8）户外（4）公共场所（1）
手机平台	电话（10）微信（2）网购平台（1）外卖平台（1）预约平台（1）
其他	中介公司（1）证券公司（1）银行（1）ATM机（1）旅行社（1）警察局（1）

表 5　生活类场景表

以上是根据话题所总结出的场景，共81个生活类场景。对于一些常见场景我们进行了更深一层的细分，如“医院”是一个大的场景，主要是“生病就医”一级话题的交际场景，“生病就医”一级话题下又有多个二级话题，那么本文在“医院”这一大场景下又细分了“诊室、病房”这些常见的小场景。另外，本文增加了手机平台场景，如“购物”话题中的网店，“外卖”话题中涉及到的外卖平台。目前网络平台在我们日常生活中占据了很重要的位置，这些平台上的对话虽不是常规的口语，但是具备口语色彩。对留学生来说，有必要学习常用的网络平台的对话进行交际。

通过对话题下场景的分析，我们发现一些场景几乎包含了所有话题，如“家、宿舍”这些固定生活场所，而大部分场景下话题受限制较大，在我们收集到的语料中一般只包含一个话题，如“医院诊室、病房、手术室外”场景一般只涉及“生病就医”话题、“音乐厅、游乐场”等场景一般只涉及“休闲娱乐”话题。这些场景内人们的对话通常是针对某一话题展开，话题延展性不强。由此可知大多话题和场景的关联性较强。

5　口语素材库在教材自动定制中的应用分析

基于上文口语话题-场景素材库的研究，本文对素材库的实现进行了前期的测试，可根据场景类型进行场景的选择，如图4:

图 4　场景选择模式测试

在场景选择模式中，用户可根据场景关键词进行学习场景的选择，图4所示场所类别选择为“医”，场景即为“生病就医”相关场景，为“药店、医院”等。当用户选择场景为“医院”时，学

习内容推送如图5：

场景	一级话题	二级话题	三级话题	句子	句子组序号	话轮中的句子序号
医院	日常交际	常用语	安慰	我这病怕是好不了了。	1	1
医院	日常交际	常用语	安慰	别担心，会好的。	2	1
医院	日常交际	常用语	劝阻	老刘的病很严重，咱们应该把病情告诉他。	1	1
医院	日常交际	常用语	劝阻	最好不要吧，他会受不了的。	2	1
医院	生病	费用	报销	你好，请问我的医保卡可以报销吗？	1	1
医院	生病	费用	报销	我看一下你的医保卡。	2	1
医院	生病	费用	报销	哦，你这个是只能住院才可以报销的，平时的买药是不能报销的。	3	1
医院	生病	费用	报销	哦，那好吧。	4	1
医院	生病	费用	报销	你知道咱们学校的医保可以报销多少吗？	5	1
医院	生病	费用	报销	据说是可以报销90%。	6	1
医院	生病	费用	报销	吓，那么多呀？	7	1
医院	生病	费用	报销	对呀，所以说在咱们校医院看病是很划算的。	8	1
医院	生病	费用	报销	我正好感冒了，我要去拿点药。	9	1
医院	生病	费用	报销	你看本来这些药，如果在药店买的话需要100多块钱，但是咱们校医院报销过以后只用花十几块钱。	10	1
医院	生病	费用	报销	咱们学校报销的真的挺多的。	11	1
医院	生病	费用	报销	对。	12	1
医院	生病	费用	缴费	好的，你拿着这个单子交一下费。	1	1
医院	生病	费用	缴费	出了门有一个自助的交费机器，你就在那上面交就可以。	2	1
医院	生病	费用	缴费	好的这是你的病例单，下一次过来的时候拿着这个和预约单。	3	1
医院	生病	费用	缴费	那我还用下面去，下面挂号吗？	4	1
医院	生病	费用	缴费	不用去了，你就直接来上面排队就可以。	5	1
医院	生病	费用	缴费	好的谢谢医生。	6	1
医院	生病	挂号		您得先填写这张挂号表。	1	1
医院	生病	挂号		我应该挂哪科呢？	2	1
医院	生病	挂号		你最好挂内科，坐电梯到3楼左拐沿着走道走，你会看到一块牌子在你右手边儿，在那	3	1

图 5　场景为“医院”的学习内容测试

在场景“医院”下，涉及到一级话题“日常交际”、“生病就医”及其下的二级、三级话题，图5所示句子组序号表示该句为所在话轮的第几句。此测试可以根据学习者的场景特定需求为其推送学习素材。

另外我们对用户的学习界面进行了设计，以学习者的就医需求为例进行口语素材推送分析，学习内容包括话题常用词、对话及常用句：

图 6　学习模块页面设计

在学习内容界面，常用词是口语素材库中归纳出的各低级话题的常用词，对常用词的学习配备相应的拼音、词语发音、词语朗读纠音以及图片展示。口语会话学习内容也保留了口语会

话的特点，在会话范读中保留口语的语音语调，但语速不宜过快。最后常用句的学习中，红色词语是可以被替换的内容，如第二句中，“头疼、嗓子疼”中的“头、嗓子”可以替换为“牙、肚子”等身体部位。常用句是对口语会话的总结，来源于上文口语素材研究中的资源。

6 结语

本文基于10341条口语语料，对适合在线教材自动推送的汉语口语素材进行了分析。本文在分词与词频统计的基础上对汉语口语词汇特征进行了深入描写，总结各词类分布特点，同时根据腾讯AL LAB 公开的中文词向量数据，使用Kmeans算法对口语词汇进行词语聚类，将全部词语分别聚类为500类、800类和1000类，通过对聚类结果的分析，发现聚类总数越多，同一话题内词汇的聚类数也越多，分类越细致，每类之间的区分度越弱，因此本文参考500类的分类结果对大话题进行划分，参考800类和1000类的结果对下级话题和场景进行归纳。其次，基于词语聚类，在对语料场景话题及现有话题库的考察下，本文构建了一个包含15个一级话题、102个二级话题以及81个交际场景的话题-场景框架体系，并对各级话题下的常用词及常用句进行了总结。本文最后对口语素材库在教材自动推送中的实现进行了前期测试，并且对用户的学习界面进行了设计，口语素材库可以实现根据学习者的场景特定需求为其推送学习素材。下一阶段的研究将针对素材库的推送进行更深层次的分析，制作具有实用价值的汉语教材自动推送产品，服务于汉语学习者与汉语教师。

参考文献

陈珏铭. 2015. 基于话题及交际图式的汉语会话常用词和常用句研究. 暨南大学.

崔彩霞. 2008. 停用词的选取对文本分类效果的影响研究. 太原师范学院学报:自然科学版,7(04):91–93.

方沁. 2013. 基于话题分类的汉语教学影视片段资源库构建. 暨南大学

官琴,邓三鸿,王昊. 2017. 中文文本聚类常用停用词表对比研究. 数据分析与知识发现,1(03):72–80.

国家汉语国际推广领导小组办公室. 2008. 国际汉语教学通用课程大纲. 北京:外语教学与研究出版社.

刘华. 2019. 基于文本分类中特征提取的领域词语聚类. 语言文字应用,2007(01):139–144.

吕荣兰. 2011. 基于语料库的对外汉语口语话题及话题词表构建. 暨南大学.

苏新春,唐诗瑶,周娟. 2011. 话题分析模块及七套海外汉语教材的话题分析. 江西科技师范学院学报,2011(06):58–65.

杨继洲. 2000. 对外汉语初级阶段功能大纲. 北京:北京语言大学出版社.

俞士汶等. 2003. 现代汉语语法信息词典详解（第二版）. 清华大学出版社.

喻雪玲. 2013. 基于语料库的商务汉语话题库及话题词表构建. 暨南大学.

郑艳群. 2012 . 多属性标注的汉语口语教学多媒体素材库建设及应用. 语言教学与研究,2012(05):34–39.

周小兵. 2010. 建设数字化国际汉语教学资源库. 华文教学与研究,2010(01):1.

Song Yan,Shi Shuming,Li Jing,Zhang Haisong. 2018. *Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings.* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),175–180..