文章编号：1006-2467(2016)07-1102-06　　DOI：10.16183/j.cnki.jsjtu.2016.07.020

利用深度去噪自编码器深度学习的指令意图理解方法

李瀚清[1,2a]，房　宁[2b]，赵群飞[1,2a]，夏泽洋[3]
(1.上海市北斗导航与位置服务重点实验室，上海 200240；
2.上海交通大学 a.自动化系，b.人文学院，上海 200240；
3.中国科学院深圳先进技术研究院，广东 深圳 518055)

摘　要：提出了一种利用深度去噪自编码器(SDAE)的自然语言指令意图理解方法.根据家庭服务机器人的使用环境和应用场景构建了一个自然语言文本指令语料库，并对语料库中各类指令进行意图标注，从而把文本指令理解问题转化为文本分类问题；在传统的文本向量空间模型的基础上，融合了文本指令的词性信息，定义了一种文本表示模型——词性向量空间模型；将 SDAE 应用于文本指令意图理解，提取指令的高阶特征；用高斯核支持向量机进行训练和预测，进而实现了自然语言指令的意图理解.在所建语料库上进行多折交叉验证，结果表明指令意图理解平均准确率达到 96%以上.

关键词：意图理解；向量空间模型；支持向量机；深度去噪自编码器
中图分类号：TP 24　　**文献标志码**：A

Deep Learning of Instruction Intention Understanding Using Stacked Denoising Autoencoder

LI Hanqing[1,2a]，*FANG Ning*[2b]，*ZHAO Qunfei*[1,2a]，*XIA Zeyang*[3]
(1. Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai 200240, China;
2. a. Department of Automation, b. School of Humanities, Shanghai Jiaotong University, Shanghai 200240, China; 3. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China)

Abstract: This paper proposed a natural language based text-instruction intention understanding method by using the stacked denoising autoencoder (SDAE). A corpus of text-instructions in natural language was created according to the working environment and application scenarios for the home service robot. Then each text instruction was labelled with a corresponding intention so as to transform the problem of intention understanding in natural language into a simple classification one. Based on the traditional text vector space model, the part-of-speech vector space model which contains the information of part-of-speech was defined. SDAE was applied to text-instruction intention understanding for extracting the higher level features of the text-instructions. Finally, the support vector machine was used for training and prediction in order to achieve text-instruction intention understanding in natural language. The result of 10-fold cross validation on the corpus shows that the average accuracy of text-instruction intention understanding rea-

收稿日期：2015-08-17
基金项目：国家自然科学基金资助项目(51305436)
作者简介：李瀚清(1991-)，男，江西省南昌市人，硕士生，研究方向为人-机器人交互，图像处理与模式识别.
E-mail：lihanqing@sjtu.edu.cn. **赵群飞(联系人)**，男，教授，E-mail：zhaoqf@sjtu.edu.cn.

ches more than 96%.

Key words: intention understanding; vector space model (VSM); support vector machine (SVM); stacked denoising autoencoder (SDAE)

随着经济社会的发展和人口老龄化的加速，对于家庭服务机器人的需求日益加大，其应用范围也越来越广泛. 高效、友好的人-机器人交互是家庭服务机器人技术中非常重要的一个环节. 在实际应用中，交互方式多种多样，主要包括：触控交互、体感交互、语音交互等. 在诸多交互方式中，通过语音，尤其是使用自然语言与机器人进行交互是最直接、最便捷的方式. 特别是对于老龄用户以及其他肢体活动不方便的用户，基于自然语言的语音交互技术显得尤为重要，它的实现主要依靠语音识别、语音合成与语义理解等技术. 因此，对自然语言的语义理解，是人机语音交互技术中非常重要的研究内容之一.

为了使机器理解人类自然语言语义，在长期研究的基础上形成了 2 种基本方法：基于规则的方法[1-3]和基于统计的方法[4-7]. 前者从语言学的角度出发，希望建立一组语言学规则，使机器可以按照这组规则来正确理解它面对的自然语言；后者从统计学的角度出发，希望通过对大规模语料库的统计学习，使机器可以理解人类的自然语言之语义.

如今进入大数据时代，基于统计学习的自然语言理解研究将有广阔的发展前景. 与基于规则的方法相比，基于统计学习的方法更加具有扩展性和非受限性. 用户能够受到最小的限制去操纵机器人，人与机器人之间的交互就如同人与人之间的交流一样自然、流畅、高效. 用户不必为了操作机器人而专门学习一套特定的控制指令，可以在交互过程中使用日常交流时的语言进行控制，避免不必要的操作负荷. Li 等[8]利用收集的描述路径的语料库，通过建立移动机器人导航意向图，提出了一种基于受限自然语言的移动机器人视觉导航算法. Banchs 等[4]通过搜集大量的对话语料库，采用了一种双重搜索策略，实现了一个可与人聊天的对话系统. Matuszek 等[9]通过建立语义分析模型，实现了从英文指令到机器人控制指令的转化，利用自然语言指令控制机器人的动作. Dave 机器人可以通过用户的自然语言指令的学习，在可变化的室内环境下完成一些特定的任务，使自己适应变化的环境[10]. 以上研究侧重于通过利用自然语言来辅助机器人更好地完成任务，涉及的指令体系比较简单.

本文设计并实现了一套基于自然语言文本指令理解的人-机器人交互系统. 通过收集家庭服务机器人的应用环境和作业任务，形成一套有效的机器人控制指令体系，并创建一个基本覆盖家庭服务的文本指令语料库. 针对语料库中大部分控制指令语句短小、口语化严重等特点，在向量空间模型(Vector Space Model, VSM)[11]的基础上，提出一种融合了词性信息的词性向量空间模型(Part-Of-Speech Vector Space Model, POS-VSM)，更充分地表现指令中的语义信息. 将图像处理领域中备受关注的深度去噪自编码器(Stacked Denoising Autoencoder, SDAE)[12]应用于意图理解，提取指令的高阶特征，使得系统具有更强的鲁棒性. 用支持向量机进行训练和预测，实现对自然语言文本指令的意图理解.

1　人-机器人交互系统框架

人-机器人交互系统框架如图 1 所示，主要包括与用户直接进行交互沟通的输入输出层(IO 层)，对用户意图进行理解、管理及应答的控制层和包含相关语料库的知识层.

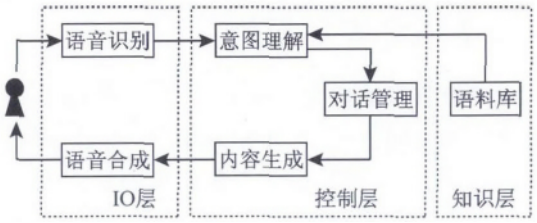


图 1　人-机器人交互系统框架

Fig. 1　Framework of human-robot interactive system

在 IO 层中，先将用户语音指令识别为相应的文本指令，传递给控制层. 另一方面将控制层中输出文本应答信息合成为语音信号，对用户进行应答.

控制层主要包括意图理解、对话管理、内容生成 3 大模块. 意图理解模块是本系统的核心模块，其根据文本指令识别用户的意图. 本文通过对知识层中的语料库进行人工标注，即对每句指令标注对应的意图，从而把意图理解问题转化为文本指令分类问题. 在意图理解模块中，按照概率大小依次输出 3 个可能的意图，传递给对话管理模块. 对话管理模块根据上下文进一步对意图进行判断，由内容生成模块生成应答文本，输出至 IO 层.

知识层存储了系统学习的语料库及预设的意

图. 如图 2 所示，语料库包含了音乐播放、邮件收发、电视控制、煮饭、电话、空调控制、天气预报及系统设置 8 大类文本指令，共 60 个意图，总计 6 000 余条，基本涵盖了家庭服务的主要内容，并用一个树形结构表示该指令体系. 例如，用户想发起一个给张三打电话的操作，可以说“给张三拨打电话”. 在该指令体系中设置了一个 4 层的意图树，层层递进，对操作进行了很详细的划分. 本文所指的意图是指意图树中的节点，将每句用户指令进行识别，识别出对应意图树中处于哪个节点，然后根据前后文进行一个对话管理，将用户引导至意图树的终点完成对应操作. 当需要扩展该对话系统以便识别更多的意图时，只需在语料库中增加相应的训练数据后重新训练，而不需要对整个系统进行修改，可以很方便地拓展到其他应用中去.

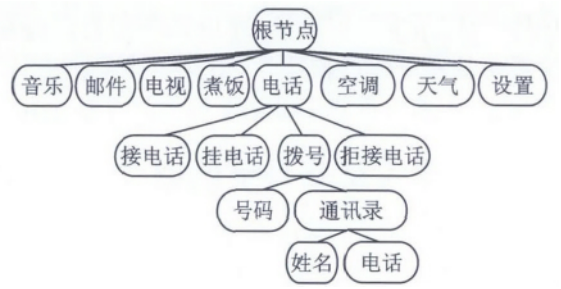


图 2 指令体系

Fig. 2 Instruction system

2 文本指令表示模型

文本指令表示模型主要有 3 种：布尔模型、概率模型和向量空间模型. 在布尔模型中，文本用标引词集合表示，优点在于具有清楚和简单的形式，但是完全匹配会导致太多或者太少的结果被返回. 在概率模型中，用一个表示语言基本单位（词、词组、句子等）的分布函数描述该语言基于统计的生成规则，如常见的概率语义分析模型（PLSA）和隐性语义分析（LDA）[13] 等. 它克服了传统信息检索中文本相似度计算方法的缺点，并且能够在海量数据中自动寻找出文字间的语义主题. 其模型也较为复杂，需要较大的计算量. 在向量空间模型中，文本指令用多维空间的向量来表示，并且它以空间上的相似度表达语义的相似度，直观易懂，简洁高效. 当文本指令被表示为向量的形式时，就可以通过计算向量之间的相似性来度量文本指令之间的相似性. 在表达足够的语义信息的前提下，所需的运算量也非常少，十分适合大规模的运算.

2.1 向量空间模型

VSM 由 Salton 等[11] 于 20 世纪 70 年代提出，并成功地应用于著名的 SMART 文本检索系统. 它是一种非常简单高效的文本模型，至今已在许多场合应用.

在 VSM 中，每一条文本指令 I 都被映射成多维向量空间中的一个点，并且用此空间中的向量 $[W_1\ W_2 \cdots\ W_n]$ 来表示（其中，W_i 为词 T_i 对应的权值，用以表现该词在该文本指令中的重要程度，$i=1,2,\cdots,n$），从而将文本指令的表示和匹配问题转化为向量空间中向量的表示和匹配问题来处理.

对于词权重的计算，经典的 tf×idf 方法考虑 2 个因素：

（1）局部因子. 词频 tf（term frequency），即词语在文本指令中出现的次数.

（2）全局因子. 逆文本频率 idf（inverse document frequency），表征每个词语在文本指令集合中分布情况的一种量化值，常用的量化方法为 $\ln(N/n_i)$. 其中：N 为文本指令集合中的文本指令数目；n_i 为包含第 i 个词（T_i）的文本指令数.

根据以上 2 个因素，可以得到权值公式：

$$W_i = \mathrm{tf}_i \ln(N/n_i)$$

其中，tf_i 为词 T_i 在文本指令 I 中的词频.

为了计算方便，再进行归一化，最后有：

$$W_i = \frac{\mathrm{tf}_i \ln(N/n_i)}{\sqrt{\sum_{i=1}^{n} (\mathrm{tf}_i \ln(N/n_i))}}$$

2.2 词性向量空间模型

一般而言，在语音交互系统中，指令比较短，而且结构也简单，但对机器人而言却依旧难以理解，传统的 VSM 提供的语义信息不够充分. 因此，本文在 VSM 基础上，融合了词性信息，提出了一个新的文本指令表示模型，称为词性向量空间模型（POS-VSM）.

在 POS-VSM 中，将每个词的词性也加入到模型中，即在模型中另外增加词性维度，并对词性进行统计量化. 权值策略与 VSM 类似，统计语料库中所有词性 p_j（$j=1,2,\cdots,m$，即共有 m 个词性）. 局部因子使用词性频率作为量化指标，即每种词性出现的次数. 为了区别不同词性，并进一步体现短语类型（如主谓短语，介宾短语等），本文将全局因子取为

$$\ln[(N+1)/n_j]$$

其中：n_j 为 N 条文本指令中，包含词性 p_j 的文本指令个数；“1”是为了避免部分词性几乎在每条指令中都出现，导致权值为零而无法体现短语类型. 例如，在指令“我＃r＃想＃v＃去＃v＃人民广场＃n”和“我＃r＃要＃v 吃饭＃v”中，增加词性 r（代词）、v

(动词)、n(名词)3 个维度，相应的权值如表 1 所示.

表 1　VSM 和 POS-VSM 中的全局因子

Tab. 1　Global weight of VSM and POS-VSM

词/词性	VSM	POS-VSM
我	ln(2/2)	ln(2/2)
想	ln(2/1)	ln(2/1)
去	ln(2/1)	ln(2/1)
人民广场	ln(2/1)	ln(2/1)
要	ln(2/1)	ln(2/1)
吃饭	ln(2/1)	ln(2/1)
r	—	ln(3/2)
v	—	ln(3/2)
n	—	ln(3/1)

3　深度去噪自编码器

SDAE 最早由 Vincent 等[12]提出，主要应用于图像处理领域，本文将该方法应用于文本指令意图理解，并取得了良好的效果.

SDAE 是将多个 DAE(Denoising Autoencoder)叠加起来的一种算法. DAE 相比于一般的自动编码器，它在训练数据中添加了随机噪声，然后从被施加噪声的训练数据中学习、重构原始数据，提高系统对噪声的鲁棒性. 如图 3 所示，在一个 DAE 中，给定一个样本 x，按概率 P_{D} 随机腐蚀部分数据得到 $\tilde{x}$，然后通过编码器 f_θ 将它映射到 y，再通过解码器 g_θ 重构 x，得到重构的结果 z，重构误差为 $L_{\mathrm{H}}(x, z)$. 将多个 DAE 叠加起来构成深度网络，即为 SDAE. 此处随机腐蚀操作可以理解为对物体的遮蔽，重构出的数据可以理解为识别出被部分遮挡的物体.

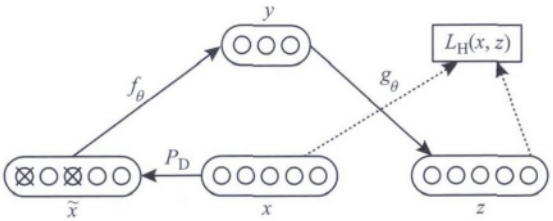
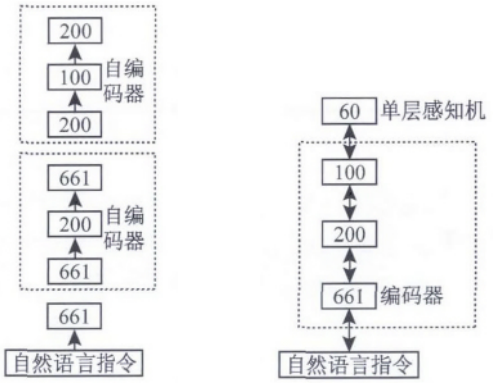


图 3　去噪自编码器结构图

Fig. 3　Architecture of denoising autoencoder

在自然语言中，尤其是口语中，助词、叹词、语气词等经常出现，却包含较少的语义信息. 这些词影响了计算机对指令的理解，如果直接删除则会丢失部分语义信息，影响识别率. 但是不加处理直接使用又容易导致过学习，影响系统的识别效果. 为解决该问题，本文采用 SDAE 去噪编码器.

首先，用 POS-VSM 模型将文本指令向量化，得到一个表示该指令的特征向量(在本文设计的系统中是一个 661 维的向量)，并将该向量输入到 SDAE 中，并对该特征向量进行随机腐蚀操作. 不同于图像处理中将该操作应用于整幅图像，本文仅在助词、叹词、语气词等维度添加随机噪声，使得训练数据更接近实际口语环境. 然后，采用无监督特征学习进行逐层训练，称为预训练，如图 4(a)所示. 在本文设计的系统当中，由 2 个 DAE 组成. POS-VSM 表示文本指令的维度为 661 维，第 1 个 DAE 隐藏层的神经元数目为 200，本文仅在该层添加随机噪声. 第 2 个 DAE 隐藏层的神经元数减到 100. 为获得更有效的网络参数，当预训练结束后，在网络的顶端连接一个单层感知机，反向传播分类误差，对整个网络进行微调，如图 4(b)所示. 微调后得到的重构数据更接近原始数据，实现了网络参数优化. 至此，本文得到了一个可以获取文本指令高阶特征的深度网络，即图 4(b)虚线框中的编码器，它以 661 维的特征向量为输入，输出一个包含文本指令高阶特征的 100 维特征向量.



(a) 预训练过程　　(b) 微调过程

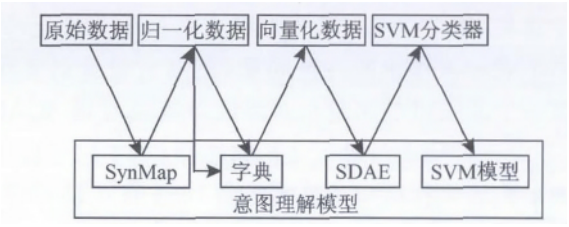
图 4　SDAE 训练过程

Fig. 4　Training process of SDAE

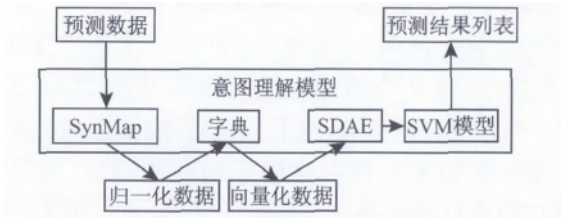
4　意图理解流程

本节将要介绍意图理解模块的整个流程，如图 5 所示. 该模块的输入是语音识别得到的文本，并经过分词处理的词序列，输出为意图列表及相应概率. 意图理解过程主要分为离线训练与在线预测两部分.

在离线训练过程中(见图 5(a))为了降低特征维度，提高识别精度，首先需要对原始数据进行预处理，对同义词进行归一化处理，即采用事先建立好的一个同义词表——SynMap，将语料库中所有同义词转化成一个统一的词，得到的数据称作归一化数据.



(a) 训练过程



(b) 预测过程

图 5 意图理解流程

Fig. 5 Process of intention understanding

然后用 POS-VSM 模型将归一化数据向量化，即用归一化的数据生成字典，并统计得到相应的全局因子，再通过该字典将语料库向量化（在本文系统中是一个 661 维的特征向量）. 随后，用向量化的训练数据库训练 SDAE 网络，提取高阶特征（在本文系统中是一个 100 维的特征向量）. 最后将 SDAE 输出的包含高阶特征的特征向量输入到 SVM 分类器中，训练 SVM 模型. 最后，得到由 SynMap、字典、SDAE 网络和 SVM 模型 4 部分组成的意图理解模型.

在在线预测过程中（见图 5(b)），类似地，将预测数据输入到意图理解模型中，通过 SynMap 将预测数据归一化，随后通过字典将其向量化（在本文系统中是一个 661 维的特征向量），再通过 SDAE 网络提取其高阶特征（在本文系统中是一个 100 维的特征向量），最后输入到 SVM 模型中，根据概率从大到小进行排序，输出最终预测结果列表. 预测结果列表只取前 3 个意图预测结果及其对应概率.

需要注意的是，一般文本分类采用的是线性 SVM 分类器. 由于经过 SDAE 网络提取的高阶特征是非线性的，本文在 SVM 中采用高斯核的非线性分类器，其效果可在以下的实验中得到验证.

5 实 验

在实验中，实验数据库采用的是本文第 2 节中搜集的家庭服务语料库，共 6 000 余条语料，包含 60 个意图. 在该语料库进行 10 折交叉验证以验证本文方法的有效性. 评价指标为意图预测结果列表中第 1 个、第 1～2 个，第 1～3 个预测结果的准确率，即第 1 个预测结果就正确的准确率、第 1～2 个预测结果包含正确结果的准确率和第 1～3 个预测结果包含正确结果的准确率，分别称之为 one-best、two-best 和 three-best 识别准确率.

本文提出的 POS-VSM 模型与 SDAE 相结合方法的十折交叉验证结果如表 2 所示. 由表 2 可见，平均识别准确率 one-best 达到 96.23%，two-best 和 three-best 分别达到了近 98%和 99%. 因此，本文提出的方法可以非常有效地对文本指令进行理解，识别其对应的意图.

表 2 POS-VSM＋SDAE＋SVM 10 折交叉验证准确率

Tab. 2 Accuracy of 10-fold cross validation of POS-VSM＋SDAE＋SVM

序号	识别准确率/%		
	one-best	two-best	three-best
1	96.14	97.72	98.77
2	96.67	98.25	99.12
3	95.44	97.54	98.60
4	95.79	97.72	98.77
5	96.67	98.07	98.95
6	95.26	96.84	97.89
7	96.67	98.60	99.47
8	96.14	97.89	98.95
9	96.32	97.89	98.95
10	97.19	98.77	99.65
平均	96.23	97.93	98.91

为了对比本文方法与已有方法的效果，依次对比了传统的 VSM 模型（VSM＋SVM，方法 1）、本文提出的 POS-VSM 模型（POS-VSM＋SVM，方法 2）以及 POS-VSM 模型与 SDAE 组合（POS-VSM＋SDAE＋SVM，方法 3）3 种方法. 其中前 2 种未采用 SDAE 的方法采用线性 SVM 分类器 liblinear[14]，第 3 种方法采用了 SDAE，且采用的是非线性 SVM 分类器 libsvm[15]，对应核函数为高斯核. 实验结果如图 6 所示. 实验结果表明，本文提出的 POS-VSM 模型相对于传统的 VSM 模型，one-best 准确率提高了约 2%，two-best 与 three-best 也有些许提高. 将 SDAE 引入后，进一步提高了识别精度，one-best 准确率达到了 96.23%，two-best 和 three-best 也有不同程度的提高. 因此，本文提出的 POS-VSM 模型相对于 VSM 模型，包含了更多的文本信息，可以提高识别准确率. 同时，SDAE 的引入，增强了系统的鲁棒性，进一步提高了识别的准确率，是一个非常有效的特征提取方法.

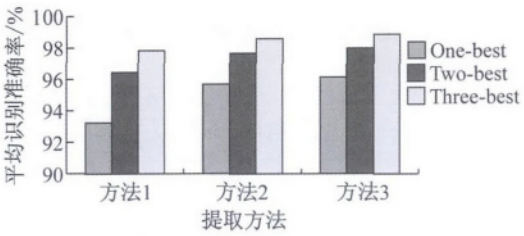


图 6　实验结果

Fig. 6　Experiment result

6　结　语

本文针对家庭服务机器人，创建了具有 6 000 余条文本指令的语料库，并设计了一套有效的面向家庭服务机器人的控制指令体系. 提出了一种融合词性信息的 POS-VSM 文本指令表示模型，能更有效地表现文本指令中的语义信息，并将图像处理领域中的深度去噪自编码器应用于基于自然语言的文本指令的意图理解，可以提取到文本指令的高阶特征，增强了系统的鲁棒性，进一步提高了意图识别准确率，one-best 意图识别准确率平均达到 96% 以上.

参考文献：

[1] CHEN Y N, WANG W Y, RUDNICKY A I. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing[C]// **Automatic Speech Recognition and Understanding**. Olomouc, Czech Republic: 2013 IEEE Workshop on IEEE, 2013:120-125.

[2] PAPPU A, RUDNICKY A. The structure and generality of spoken route instructions[C]// **Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012: 99-107.

[3] CRUTZEN R, PETERS G J Y, PORTUGAL S D, *et al*. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study[J]. **Journal of Adolescent Health**, 2011, 48(5): 514-519.

[4] BANCHS R E, LI H. IRIS: A chat-oriented dialogue system based on the vector space model[C]// **Proceedings of the ACL 2012 System Demonstrations**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012: 37-42.

[5] KOLLAR T, TELLEX S, ROY D, *et al*. Grounding verbs of motion in natural language commands to robots[C]// **International Symposium on Experimental Robotics**. Morocco: Springer Berlin Heidelberg, 2014:31-47.

[6] YAO K, ZWEIG G, HWANG M Y, *et al*. Recurrent neural networks for language understanding[C]// **INTERSPEECH**. Lyon, France: IEEE Signal Processing Society, 2013: 2524-2528.

[7] **许元辰. 基于优化的语义理解与 SVM 相结合的文本情感分类研究[D]. 南昌：南昌大学信息工程学院，** 2014.

[8] LI X D, ZHANG X L, DAI X Z. A visual navigation method of mobile robot based on constrained natural language processing[J]. **Robot**, 2011, 33(6): 742-749.

[9] MATUSZEK C, HERBST E, ZETTLEMOYER, *et al*. Learning to parse natural language commands to a robot control system[C]// **International Symposium on Experimental Robotics**. Singapore: Springer International Publishing, 2013:403-415.

[10] MISRA D K, SUNG J, LEE K, *et al*. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions[J]. **The International Journal of Robotics Research**, 2016, 35(1/2/3): 281-300.

[11] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. **Communications of the ACM**, 1975, 18(11): 613-620.

[12] VINCENT P, LAROCHELLE H, LAJOIE I, *et al*. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. **The Journal of Machine Learning Research**, 2010(11): 3371-3408.

[13] LU Y, MEI Q Z, ZHAI C X. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA[J]. **Information Retrieval**, 2011, 14(2): 178-203.

[14] FAN R E, CHANG K W, HSIEH C J, *et al*. LIBLINEAR: A library for large linear classification[J]. **The Journal of Machine Learning Research**, 2008, 9: 1871-1874.

[15] CHANG C C, LIN C J. LIBSVM: A library for support vector machines[J]. **ACM Transactions on Intelligent Systems and Technology (TIST)**, 2011, 2(3): Article 27.