|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **社会计算结课论文** | | | | |
|  | | | | |
| 人机对话系统中用户意图领域分类方法研究 | | | | |
|  | 姓名 | ： | 钱志浩 |  |
| 学号 | ： | 2018302170027 |
| 专业名称 | ： | 信息安全 |
|  |  |  |
|  | | | | |
| 二〇一九年一月九日 | | | | |

# Abstract

As a branch of artificial intelligence, the human-computer dialogue systems have important research significance and practical value. This paper analyzes the components of the dialogue systems and emphasizes the importance of the classification of user intention domain. Based on the overview of the current research situation, several approaches to the classification of user intent domain are compared.

In this paper, an algorithm based on the traditional intention domain classification method is proposed. Based on the improved word2vec hidden layer function based on the experimental content, word vectors are constructed, and softmax regression algorithm is used to select the final classification.

Keyword: human-computer dialogue systems; classification of intent domain; word2vec; softmax

# 摘要

人机对话系统作为人工智能领域的一个分支有其重要的研究意义和实用价值。本文分析了人机对话系统的组成部分，重点强调了用户意图领域分类的重要性。通过对当前研究状况的概述，对比了处理用户意图领域分类问题的多种方法。

本文提出了一种基于传统意图领域分类方法的算法，在根据实验内容改进word2vec隐含层函数的基础上构造词向量，运用softmax回归算法实现最终对分类的选择。

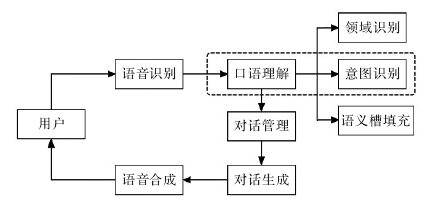
关键词：人机对话系统；意图领域分类；word2vec；softmax

# 引言

随着智能时代的推进，人工智能在现代生活的方方面面都有其应用，包括图像识别、语音识别、自然语言处理等。淘宝智能客服、天猫精灵AI音响等这样一些具有代表性的人工智能产品，其核心是人工智能的一个研究领域——人机对话系统。社会对这样的智能产品有着广泛的需求，它们可以免去很多重复的问答活动。将回答的一方给到人工智能，在保证效率的同时，还能节省更多人力。

而构建人机对话系统，是一个综合性的行为，包括语音识别、口语理解、对话管理、对话生成、语音合成这五个部分[1]，如图1所示。其中，最为关键的一个环节是口语理解。在人机对话系统的应用过程中，用户可能会有多种意图，相应地会触发人机对话系统中的多个领域，如查询机票、酒店、公交车等。人机对话系统的一个关键任务就是正确地将用户的意图分类到相应的领域中，从而才能在相应的领域中检索相关问题的答案，进一步给出正确的答复。不止于此，准确地分类用户意图还有助于智能系统给用户提供更有针对性的建议和帮助，提高用户体验。

因此，对用户意图领域分类的研究具有重要的实用价值。

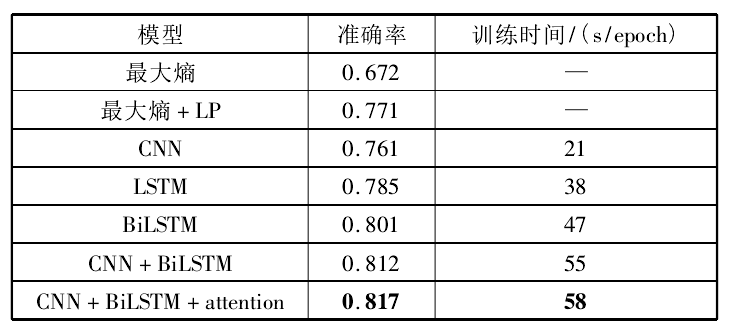
图 1 对话系统示意图

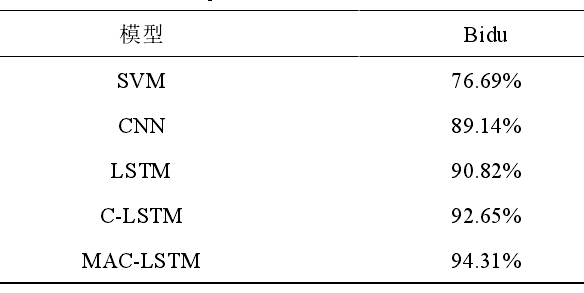
# 研究现状

就实现过程而言，对用户意图领域的分类可分为两个步骤：第一，关键词提取；第二，领域匹配。难点在于，关键词可能涵盖多个领域，比如关键词“火车”，可能是问火车票的订购，也可能是对火车时刻表的查询，这对领域匹配造成很大的不确定性。再比如“明天去上海开会”，可能是在问机票的订购，这一类属于隐含的意图，并不能通过提取关键词得出，阻碍了进一步的领域匹配。

传统的意图领域分类方法有基于统计特征分类的方法。基本思路为对文本进行关键词的提取，如One-hot、N-Gram、Word2vec、LDA[2]、BTM[3]等，然后采用机器学习的方法进行分类。常用的方法有朴素贝叶斯[4]（Naive Bayes，NB）、支持向量机[5]（Support Vector Machines，SVM）、逻辑回归（Logistics Regression，LR）等。但是这些基于原始文本语义的关键词提取不能很好地解决隐含义的问题，而且还面临向量矩阵稀疏的问题。如陈浩辰[5]用SVM算法对微博中的消费意图进行挖掘，其分类准确率虽然达到了75%左右，但是他也发现对有隐含义的微博处理效果较差，特征提取的准确率不是很理想，而且会导致数据稀疏问题。

相比之下，端到端的深度学习模型在意图分析上表现更优秀，也是目前意图分类研究的主流方式[6]。深度学习模型可以根据给定的训练文本自动学习类别特征，能够解决传统方法所不能发现的隐含义问题，同时降低了特征提取的工作量，提高了特征提取的精确度。常用的方法有卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Networks，RNN）、长短时记忆（Long Short-Term Memory，LSTM）、门控循环单元（Gated Recurrent Unit，GRU）等。文献[7]和文献[8]都讨论了多种方法，并进行了性能比较。由图2和图3，我们可以很直观地感受到深度学习的优势。

图 2 文献7比较结果

图 3 文献8比较结果

# 方法

本次实验，笔者采用的方法是先对原始文本进行关键词提取，对于不同的分类建立不同的语料库。然后，采用改进的word2vec的方法获取词向量。

在对分类算法的选择上，笔者选用了softmax回归算法。

## 基础理论

### Word2vec词向量模型

传统的word2vec主要分为CBOW和Skip-Gram两种模式，前者是从上下文预测当前词，后者则通过当前词来预测上下文[9]。笔者根据选题做出了改进，并没有采取上下文预测的方式，而是对于输入的短句，首先对所有话题域分类的语料库进行one-hot处理，然后分别乘以对应的词频，累加得对于每个分类的评价分数。所有分类的评价分数又构成了最后的31维词向量，每一维对应该分类的评价分数。公式为：

其中，Vector[i]表示词向量的第i维，即第i个话题域的评价分数，wi表示话题域中关键词的总数，one\_hot[i,j]表示对于第i个话题域该短句的初始one\_hot值，word\_count[i,j]表示第i个话题域的第j个关键词的词频。



### Softmax

Softmax的归一性在多分类问题中有着很清晰的选择导向，因而笔者选用了softmax对词向量进行模型训练，最终选择softmax值大的分类作为结果。Softmax的原理是将对不同分类的评分映射到概率上，更为直白地表现出分类判断的倾向。假设有一个数组V，Vi表示V中第i个元素，那么这个元素的softmax值为：

## 我的方法



### 输入

笔者的模型采用了笔者自己训练的词向量作为输入，然后对词向量进行softmax回归处理，得到最终的分类概率。

### 输出

笔者的模型依据分类概率，选择概率最大的分类作为输出。

# 实验与结论

实验训练集大小为2299个短句，涉及30个垂类领域（如app，bus等）以及一个chat闲聊类，验证集大小为697个短句。

## 对比实验及结果分析

笔者对语料库建立时的筛选词频选取、分词时的停用词表（标点及无意义的虚词如怎么、什么）使用与否以及隐含层中用于处理one\_hot的词频选取方式进行了基于控制变量法的对比实验。原始实验选择了筛选词频freq=0，使用停用词表和隐含层词频选取方式为（每个话题域语料库中每个关键词的）出现次数/。对比实验结果如表1所示：

|  |  |  |
| --- | --- | --- |
| **实验变量** | **选择** | **准确率** |
| 隐含层词频选取 | 出现次数/总关键词数 | 0.7390 |
| 出现次数/ | 0.8013 |
| 出现次数/总关键词数 | 0.6338 |
| 停用词表 | 使用 | 0.8013 |
| 不使用 | 0.7688 |
| 筛选词频选取 | Freq = 0 | 0.8013 |
| Freq = 1 | 0.7948 |
| Freq = 2 | 0.7649 |

表 1 对比实验结果

实验结果证明，出现次数/的隐含层词频选取更能突出关键词与话题域之间的联系，对于类似话题域的细分有着更优秀的表现。而停用词表的使用可以排除一些无用的高频词，突出具有实际意义的关键词。筛选词频则是选择0最优，因为语料库越大，对于该领域的特征表述更准确，所以评价分数也会相应越高。

## 错误分析

在查看错误的时候，发现类似于“从东莞去北海的汽车”判定成了flight，“去肇庆市怎么坐车？”判定成了map，而这两个原本都应该是属于bus。应该是“从”、“到”、“去”这样的词属于的话题域太多，不具有代表性，不应该视为关键词，但是在停用词表中加入这些词后，反而导致了准确率的下降。

# 未来展望

本次实验尚有许多不足，比如没有处理隐含义，筛选词频为0导致矩阵较稀疏，不知道如何改进错误等问题。本次实验也有可惜的地方，如果数据集再大一点，表现会更为优秀。未来笔者有待在深入学习方面加强，希望能最终运用BiLSTM、GRU等神经网络算法完善对多分类问题的处理方式。

# 6 参考文献

[1]刘娇,李艳玲,林民.人机对话系统中意图识别方法综述[J].计算机工程与应用,2019,55(12):1-7+43.

[2]郭庆. 基于图与LDA的中文文本关键词提取算法[D].北京邮电大学,2019.

[3]黄贤英, 谢晋, 龙姝言. 融合词向量及BTM模型的问题分类方法[J]. 计算机工程与设计, 2019, 40(02):91-95.

[4]Martinez F F, Ferreiros J, Cordoba R, et al. A Bayesian NETWORKS approach for dialog modeling: The fusion BN[C]// 2009.

[5]陈浩辰. 基于微博的消费意图挖掘[D].哈尔滨工业大学,2014.

[6]黄佳伟. 人机对话系统中用户意图分类方法研究[D].华中师范大学,2018.

[7]陈柯锦, 许光銮, 郭智, et al. 基于混合神经网络的问题分类方法[J]. 计算机与现代化, 2018(9):1-4.

[8]余本功,许庆堂,张培行.基于MAC-LSTM的问题分类研究[J/OL].计算机应用研究:1-6[2020-01-10].https://doi.org/10.19734/j.issn.1001-3695.2018.05.0452.

[9]周练.Word2vec的工作原理及应用探究[J].科技情报开发与经济,2015,25(02):145-148.