|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **社会计算结课论文** | | | | |
|  | | | | |
| 基于字符级别的卷积神经网络意图领域分类 | | | | |
|  | 姓名 | ： | 马陈军 |  |
| 学号 | ： | 2018302180149 |
| 专业名称 | ： | 信息安全 |
|  |  |  |
|  | | | | |
| 二〇一九年一月十日 | | | | |

# Abstract

In recent years, with the development of the artificial intelligence, man-machine dialogue systems that can greatly improve the efficiency of human-computer communication have received widespread attention from the academic and industrial communities. In the application process of man-machine dialogue system, there may be multiple user intents that will trigger the accordingly many fields in man-machine dialogue system,which includes task-vertical Field (such as query airfare, hotel, bus, etc.), knowledge-quiz field, chatting field and other fields. Therefore, how to improve the accuracy of user intent classification is an important research work in the natural language understanding module of human-computer dialogue.

This paper studies and analyzes traditional machine learning classification methods, deep learning models, and proposes character-level intent domain classification methods based on Convolutional Neural Networks. In the process of implementation, a variety of deep learning optimization techniques were simultaneously tried, and a brief introduction and experimental analysis of these techniques

Keyword: Man-machine dialogue; intention classification; CNN

# 摘要

近年来,随着人工智能热潮的兴起,能极大地提高在提高人与计算机沟通效率的人机对话系统受到了学术界和产业界的广泛关注。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，其中包括任务型垂直领域（如查询机票、酒店、公交车等）、知识型问答以及闲聊等。因而，如何提高用户意图分类的准确率是人机对话的自然语言理解模块中一个重要的研究工作。

本文研究和分析了传统机器学习分类方法，深度学习模型，提出字符级别的基于卷积神经网络的意图领域分类方法。在实现的过程中同时尝试了多种深度 学习优化技术,并对这些技术进行了简要的介绍及实验分析。

关键词：人机对话；意图分类；卷积神经网络

# 引言

随着互联网的飞速发展，人机对话技术成为学术界和工业界的研究热点。对于学术界而言，人机对话技术对语音识别、自然语言处理、信息检索以及情感分析等领域有着重要的推动作用；对于工业界而言，人机对话技术是个人事务助理、虚拟情感陪护机器人以及娱乐性聊天机器人等产品的核心技术，其性能直接影响了产品的市场竞争力。

自然语言处理（NLP）是人工智能的重要组成部分，而人机对话则是自然语言处理能力的集中表现。这其中涉及了多个 NLP 中的关键技术，包括语音识别，意图识别，句法分析，语义分析，知识库等。在本次实验的用户意图领域分类任务中，我们需要对用户的意图进行领域分类，类别包含闲聊（chat）和垂类两大类，其中垂类又细分为 30 个垂直领域（如 app,bus,calc 等），即判断人机对话系统中用户的一个输入（一段文本）所属的类别。意图识别是整个过程的第一步，给用户对话一个宏观意图领域上的分类，如用户想查询机票、咨询客服问题、还是闲聊等，为后面的用户具体目的识别缩小了判断范围，减少任务难度，而只有将用户输入分到正确的意图领域内，后面的处理才有可能正确，最终返回适当的结果。因此，意图识别在人机对话系统中占有重要的位置。一般情况下可以将意图识别分为意图领域分类、意图分类、关键信息抽取三个子任务。

意图领域分类，从根本上可以认为是基于文本进行的文本分类。在人机对话中文本分类有又自己的特点，首先文本偏向于口语，其次文本较短。20 世纪 90 年代逐渐成熟的基于机器学习的文本分类方法,在分类效果和灵活性上都比之前基于知识工程和专家系统的文本分类模式有所突破,成为相关领域研究和应用的经典范例。本文主要研究了深度学习在文本分类方面的应用，使用卷积神经网络对句子建模向量表示并分类，考虑到提供的数据集规模较小且多为短文本，而同时存在类别间的数据不平衡和缺乏上下文信息等现象，故在 CNN 模型的基础上提出采用了字符级别向量构建来缓解上述问题造成的影响。

# 研究现状

传统的意图分类研究模式一般是通过机器学习方法，将标记好意图类别的对话文本作为有监督的数据来训练分类器，对用户意图进行分类。很多常规的机器学习分类方法都被应用到了意图分类中。从朴素贝叶斯到支持向量机(svm)，隐马尔可夫模型(HMM)，决策树(DT)，到基于规则的分类方法，研究者都做出了尝试，并取得了一定的效果。M Karahan结合了三种传统机器学习分类模型，采用voting方法来提高意图分类准确率。A Celikyilmaz提出一种基于LDA主题模型的方法识别意图特征词汇，提高了电影领域用户意图分类的效果。在这些传统的方法中，核心工作是提取对话文本中人工定义的语义特征如：one-hot，TF-IDF，N-gram，词典和知识，特殊的数核函数等，然后采用机器学习的方法进行分类。N-gram特征窗口长度是有限的，再加上对话的口语化语料一般较短，甚至有时候都不符合语法，这些基于原始词义提取人工特征的文本分类模型不能很好地提取对话文本的深层语义信息，而且还面临着数据稀疏的问题，特别是在应用到对话系统中意图分类这种对话短文本分类任务时，特征向量矩阵更为稀疏，因此基于传统机器学习意图分类方法的准确性一直都难以提高。由于对话系统意图分类问题的复杂性，一般研究者都专注于特定领域对话系统的意图分类如：餐厅点餐意图、酒店预定意图、出行消费意图，电商智能客服系统中用户购买意向、电影领域相关意图[1]。

深度学习模型在意图分类任务上越来越流行，目前已成为意图分类研究的主流方法。特征提取是影响意图分类好坏的关键性因素之一。传统方法的特征表示和特征提取方法费时，需要较多的领域知识，并且不能很好地提取对话文本的深层语义信息。深度学习模型可以根据给定的训练文本自动学习语义信息和类别特征，在降低特征提取工作量的同时，抓住了传统方法无法提取的特征，提高了分类的精度。以CNN(循环神经网络)、RNN(循环神经网络)，LSTM(长短记忆模型)、GRU(门循环单元)为代表的神经网络模型能够很好地表示句子或者文档的语义信息，从而在意图分类领域中得到了广泛的应用。通过研究发现词向量维度，学习率，dropout，隐藏层参数数量对分类效果有较大影响，并在多个实验数据集上训练了一个准确率较高的LSTM意图分类模型。M H Lian Meng提出了一种高维的LSTM模型HLSTM，分别采用词级别和句子级别的LSTM构成高维模型在电商智能客服系统中用户购买意图分类语料集上进行实验，取得了较好的实验效果。还提出了一种新的记忆单元，效果要优于传统的LSTM模型。X Ding基于CNN卷积神经网络提出了CIMM模型从社交网络语料中挖掘用户的商业消费意图。但单独的LSTM模型对文本局部语义特征和类别特征提取能力不强，而CNN模型对文本时序特征的刻画能力不高。

为克服单一模型的上述问题，深度学习组合模型在意图分类任务中的应用得到关注。钱岳，丁效，刘挺等人结合了CNN和LSTM的优点，构建了一种基于卷积的长短期记忆网络模型，并在出行消费意图识别任务上有一定效果。

目前的深度学习模型普遍地采用embedding技术来进行特征表示。S Ravuri分析了意图识别任务中采用embedding技术的效果。词向量或字符向量由word2vec和glove等工具在大规模文本数据上训练生成，不仅可以将词向量表示映射到低维向量空间(一般常用是50-300维)来解决传统语言模型数据维度过高问题，而且包含了语义信息。Embedding技术的出现，神经网络模型得以很好地应用到意图分类任务中。[2]

]

# 方法

## 文本向量方法

### one-hot

one-hot是最简单的向量表示方法。该方法将词表中的每一个词都表示成一个向量，每个词都用向量互不相交的维度表示，即该向量的维度就是词表的大小，除了表示当前词这一维度的值为1，其他各个维度的值均为0。例如：

“中国”表示为：[0…1,0,0,0…0,0]

“美国”表示为：[0…0,1,0,0…0,0]

但是这样的表示方法会导致任意两个词之间都是独立的，关系相近的词甚至同义词。例如：“中国”和“美国”虽然基本上同属于国家类别，但是从向量的取值上不能判断其存在词义上的关系，从而在机器学习模型中被看作完全没有关系的两个词，失去了词汇之间的语义关联信息。而且由于词表规模的庞大,会造成词向量的维度爆炸，生成的词向量矩阵是一个高维的稀疏矩阵,给机器学习模型的训练提高了难度。

### TF-IDF

TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TFIDF实际上是：TF \* IDF，TF词频(Term Frequency)，IDF逆向文件频率(Inverse Document Frequency)。TF表示词条在文档d中出现的频率。IDF的主要思想是：如果包含词条t的文档越少，也就是n越小，IDF越大，则说明词条t具有很好的类别区分能力。如果某一类文档C中包含词条t的文档数为m，而其它类包含t的文档总数为k，显然所有包含t的文档数n=m+k，当m大的时候，n也大，按照IDF公式得到的IDF的值会小，就说明该词条t类别区分能力不强。但是实际上，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，并选来作为该类文本的特征词以区别与其它类文档。这就是IDF的不足之处. 在一份给定的文件里，词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(term count)的归一化，以防止它偏向长的文件。词频-逆文档频率权重向量的计算公式如下：



其中 TF 表示词频，|D|表示文档的总个数，|{d∈D: t∈d }|表示词在多少个文档中出现过。

## 传统机器学习方法

### svm支持向量机

SVM是从线性可分情况下的最优分类面发展而来的，其基本思想可用图1所示的二维情况说明。

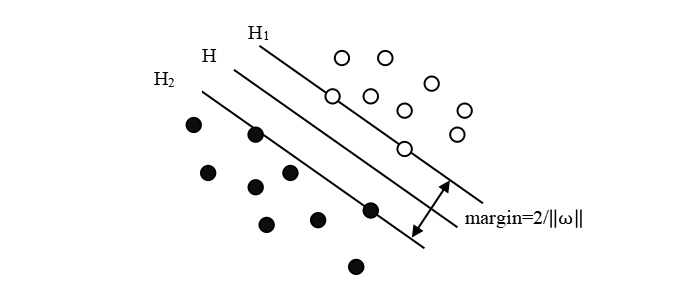


图1：最优分类面示意图

空心点和实心点代表两类数据样本，H为分类线，H1、H2分别为过各类中离分类线最近的数据样本且平行于分类线的直线，他们之间的距离叫做分类间隔（margin）。所谓最优分类线，就是要求分类线不但能将两类正确分开，使训练错误率为0，而且还要使分类间隔最大。前者保证分类风险最小；后者（即：分类间隔最大）使推广性的界中的置信范围最小，从而时真实风险最小。推广到高维空间，最优分类线就成为了最优分类面。

SVM的核函数方法就是用非线性变换将n维矢量空间中的随机矢量x映射到高维特征空间，在高维特征空间中设计线性学习算法，若其中各坐标分量间相互作用仅限于内积，则不需要非线性变换的具体形式，只要用满足Mercer条件的核函数替换线性算法中的内积，就能得到原输入空间中对应的非线性算法。

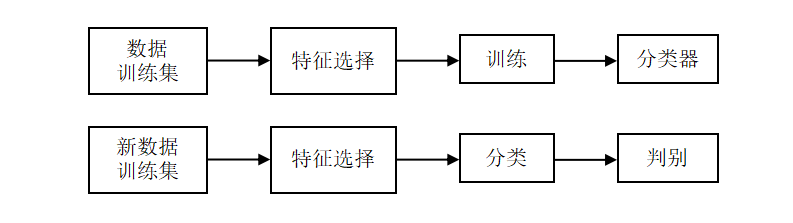
具体svm分类的工作流程如下：

图2：SVM工作流程

## 本文使用方法

### 文本预处理

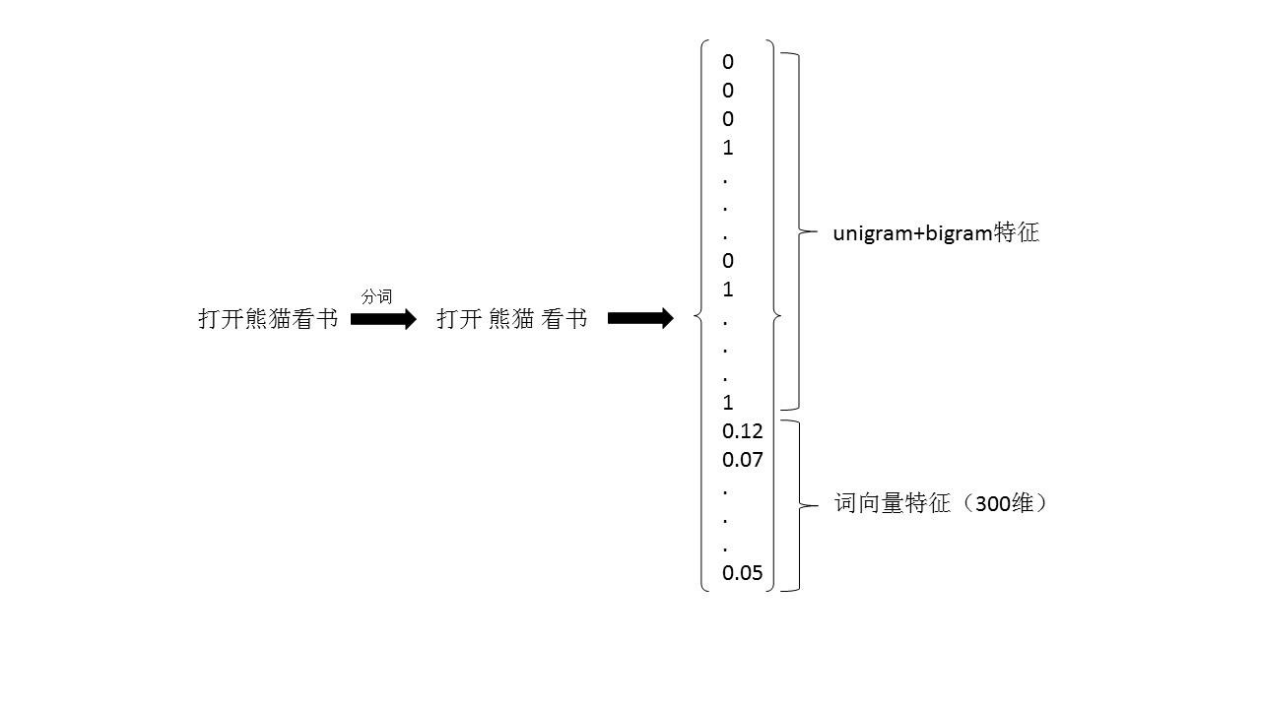
### 本次评测提供的数据共包含 31 个领域，具体包括聊天类（chat）和垂类（30 个垂直领域），垂直领域有 app、bus、calc、cinemas、contacts 等。数据分为训练集和开发集，训练集包含 2299 个文档，开发集包含 770 个文档，总共 3069 个。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CAT** | **Train** | **Dev** | **CAT(** | **Train** | **Dev** | **CAT** | **Train** | **Dev** |
| chat | 455 | 154 | health | 55 | 19 | schedule | 29 | 9 |
| app | 53 | 18 | lottery | 24 | 8 | stock | 71 | 24 |
| bus | 24 | 8 | map | 68 | 23 | telephone | 63 | 21 |
| calc | 24 | 8 | match | 24 | 8 | train | 70 | 24 |
| cinemas | 24 | 8 | message | 63 | 21 | translation | 61 | 21 |
| contacts | 30 | 10 | music | 66 | 22 | tvchannel | 71 | 23 |
| cookbook | 269 | 89 | news | 58 | 20 | video | 182 | 60 |
| datetime | 18 | 6 | novel | 24 | 8 | weather | 66 | 22 |
| email | 24 | 8 | poetry | 102 | 34 | website | 54 | 18 |
| epg | 107 | 36 | radio | 24 | 8 |  |  |  |
| flight | 62 | 21 | riddle | 34 | 11 |  |  |  |

表1：数据集不同句子类别统计

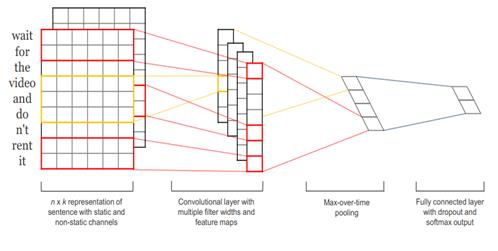
可见文本类别较多，但提供的数据规模较小，且存在数据不平衡现象，训练数据最多的chat 领域是较少的 datetime 领域数据规模的 25 倍。除此之外，经人工分析数据发现，不同领域数据间可能存在较高的相似性，如 tvchannel 领域和 epg 领域，较难通过简单的关键词匹配或线性模型分类。

由于本次实验提供的是分离的文本文件，具体构建时不易转化为词向量，因而使用了Python的文件读取写入功能，将文本按照的标签字典序写入一个文本文件中。为了便于便于后期的模型训练，本次采用词向量对文本进行特征提取。鉴于传统的one-hot,TF-IDF方法缺少词的上下文信息，因而本次采用的是word2vec工具进行向量化。word2vec是Google公司于2013年发布的一款开源的Deep Learning开发工具。它需要首先在已分词的文本语料库中构建一个词汇表。根据Skip-Gram模型或是CBOW模型通过训练得到每个词的词向量，其相似词汇之间的向量是相近的，并可以通过计算两个词向量之间的余弦相似度判断其语义相似度。通过word2vec可以将对文本内容的处理简化为向量空间中的向量运算。句子“打开熊猫看书”的基于词向量的特征具体表示如图3所示：

图3：句子“打开熊猫看书”的特征表示

### 训练模型构建

得到文本向量后，可以开始着手进行模型构建。相比与传统的支持向量机(SVM)模型，卷积神经网络(CNN)模型在文本分类任务上具有独特的优势。卷积神经网络(CNN)是一个深层神经网络结构，分为特征提取层和采样层即卷积层和池化层，具有共享权值和局部信息感知的特点。卷积神经网络神经元共享网络层权值从而减少了网络自由参数的个数从而降低学习复杂度，不同卷积核的卷积过程可以提取局部语义特征(自然语言领域中)。CNN模型在各类语料集中表现出了很好的效果，并且在少数几个语料集中获得最先进的精确度。其基本结构如图4所示。

图4 简单句子的 CNN 模型

一般认为文本分类不涉及很深层的语义，往往更关注局部特征的识别，CNN 模型在文本分类任务上能比SVM模型取得更好性能。传统方法需要人工不断尝试，来选择 n-grams 特征，而 CNN 模型的 featuremaps 可视为自动学习出重要的 n-gram 组合。在本次实验中，通过featuremaps 计算得到的特征然后对 31 个领域做分类。

### 模型创新与优化

基于以上的方法，对本次实验进行了模型构建，而在后续相关实验中，我又针对文本向量的生成与CNN模型进行了优化。

对于文本向量的生成，通过通过研究相关文献和实验对比，决定采用字符级向量表示为模型输入，而非传统的词级。因为对于本次实验，其训练集均为短文本，而进行分字能更好的保留文本本身的特征。具体影响会在后续具体说明。

对与CNN模型，本次训练集数量过少，为了防止模型可能会出行拟合现象，在本次实验中使用了dropout 技术，在模型训练时随机让网络某些隐含层节点参数不工作，测试时再让其持续有效并乘以相应缩放因子，这样既防止了训练时的过拟合，又没有直接减小网络规模。同样地，具体影响的也会在后续说明。

# 实验与结论

### 分词与分字

分词是中文文本预处理的常用方法，本节考察分词和分字对结果的影响，这里分词采用的是jieba库所提供的分词功能，而分字则是自己写的函数。最后在dropout=0.5，step=3000的情况下进行的实验，具体结果如下：

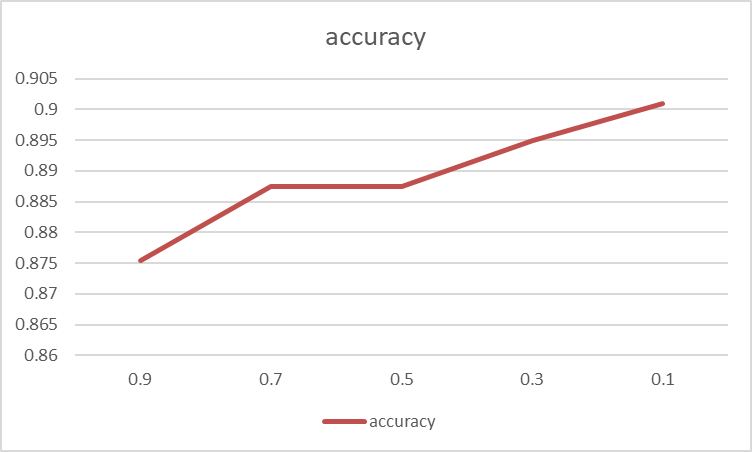
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | accuracy | loss | precision | recall | f1-score |
| 分词 | 0.8093 | 0.7583 | 0.8388 | 0.8093 | 0.8101 |
| 分字 | 0.8874 | 0.4169 | 0.8949 | 0.8874 | 0.8860 |

表2分词/分字结果表

通过表格我们可以清楚的看出，虽然分词为中文文本预处理中常用的步骤，但采用分字法可以极大的优化各类指标。猜测这是可能由于分词的过程中产生了较多的分词错误造成的，这对人机对话类的短文本影响较大，若短文本中的类别关键词出现分词错误，则直接会导致短文本的分类错误。同时也有可能是相比较于分词，分字法提高短文本的特征数量。在后续的实验部分均采用分字方法。

### dropout技术

在短文本处理中使用了dropout 技术，可以有效的减少过拟合现象，对提高分类效果有极大的意义，下图采用不同的dropout率，对准确率的影响。

图5 dropout率对实验的影响

一般来说，取0.5为比较合适，但由于本次实验训练集多为短文本且文本数量较少，故取更低也是可以的。

### 最终结果

最后选择了一组最合适的参数进行了实验，具体数据结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **category** | **precision** | **recall** | **f1-score** | **support** |
| **app** | 0.7059 | 0.6667 | 0.6857 | 18 |
| **bus** | 0.875 | 1 | 0.9333 | 7 |
| **calc** | 1 | 1 | 1 | 8 |
| **chat** | 0.8197 | 0.9804 | 0.8929 | 51 |
| **cinemas** | 0.75 | 0.75 | 0.75 | 8 |
| **contacts** | 1 | 0.9 | 0.9474 | 10 |
| **cookbook** | 0.9457 | 0.9667 | 0.956 | 90 |
| **datetime** | 1 | 0.6667 | 0.8 | 6 |
| **email** | 1 | 1 | 1 | 8 |
| **epg** | 0.8529 | 0.8056 | 0.8286 | 36 |
| **flight** | 1 | 1 | 1 | 21 |
| **health** | 0.9231 | 0.6667 | 0.7742 | 18 |
| **lottery** | 1 | 0.875 | 0.9333 | 8 |
| **map** | 0.9545 | 0.913 | 0.9333 | 23 |
| **match** | 1 | 1 | 1 | 8 |
| **message** | 1 | 0.9524 | 0.9756 | 21 |
| **music** | 0.9375 | 0.6818 | 0.7895 | 22 |
| **news** | 0.8636 | 1 | 0.9268 | 19 |
| **novel** | 0.875 | 0.875 | 0.875 | 8 |
| **poetry** | 0.8649 | 0.9412 | 0.9014 | 34 |
| **radio** | 0.8889 | 1 | 0.9412 | 8 |
| **riddle** | 1 | 1 | 1 | 11 |
| **schedule** | 1 | 0.8 | 0.8889 | 10 |
| **stock** | 1 | 0.8333 | 0.9091 | 24 |
| **telephone** | 1 | 1 | 1 | 21 |
| **train** | 1 | 1 | 1 | 23 |
| **translation** | 1 | 1 | 1 | 20 |
| **tvchannel** | 0.7619 | 0.6667 | 0.7111 | 24 |
| **video** | 0.7432 | 0.9016 | 0.8148 | 61 |
| **weather** | 1 | 1 | 1 | 22 |
| **website** | 1 | 0.7778 | 0.875 | 18 |
| **avg/total** | **0.9072** | **0.9009** | **0.8998** | **666** |

表3：最终结果

可以看出最后的成绩还是比较令人满意的。

### 结论

本实验使用基于字符级别的卷积神经网络进行用户意图领域分类。本次实验对今后的文本分类等学习打下了基础。在相关的网络比赛中，有的参赛队伍取得了很高的F值，说明了本实验仍然存在较大的提升空间，日后会继续对模型进行改进。同时我也意识到深度学习是一类强力的工具。然而在解决实际问题中，仅依赖深度学习是远远不够的。只有基于对问题的深刻理解，将深度学习与传统方法、与场景相结合，是进一步提升表现的正确途径。

# 未来展望

未来可能会考虑，使用多种模型进行训练，集成各个模型的优点，来提高文本分类的准确率。同时针对这次训练集规模较小的问题，后续也可能会考虑使用爬虫技术来扩充相关的数据集。

# 参考文献

1. 黄佳伟. 人机对话系统中用户意图分类方法研究[D]. 2018.

[2] 姜超. 基于语义的用户意图领域多分类算法分析[D].