**面向互联网文本的语义知识智能分析**

**需求分析**

题目名称：面向互联网文本的语义知识智能分析

班 级：软件162班

学生姓名：赵向阳、陈金新

指导教师：刘小明

2019 年 5 月

目录

[第1章 引 言 2](#_Toc8594993)

[1.1 编写目的 2](#_Toc8594994)

[1.1 项目背景 2](#_Toc8594995)

[第2章 任务概述 3](#_Toc8594996)

[2.1 功能需求 3](#_Toc8594997)

[2.2参与者列表 3](#_Toc8594998)

[2.2 模块管理 3](#_Toc8594999)

[2.2.1数据预处理模块管理 3](#_Toc8595000)

[2.2.2 Seq2Seq建模模块管理 3](#_Toc8595001)

[3.2.3 文本生成模块管理 3](#_Toc8595002)

[第3章 需求固定 4](#_Toc8595003)

[3.1模块需求分析 4](#_Toc8595004)

[3.1.1数据预处理 4](#_Toc8595005)

[3.1.2 Seq2Seq建模 4](#_Toc8595006)

[3.1.2 文本生成 4](#_Toc8595007)

[3.2具体用例描述 4](#_Toc8595008)

[3.3系统流程分析 5](#_Toc8595009)

[3.3.1系统主要流程 5](#_Toc8595010)

[3.3.2系统评测流程图： 6](#_Toc8595011)

[3.4非功能需求 6](#_Toc8595012)

[3.4.1访问控制 6](#_Toc8595013)

[3.4.2信息约束 6](#_Toc8595014)

[3.4.3性能 7](#_Toc8595015)

[第4章 运行环境规定 7](#_Toc8595016)

[4.1软件 7](#_Toc8595017)

[4.2硬件设备 7](#_Toc8595018)

[4.3网络 8](#_Toc8595019)

[4.4产品的质量需求 8](#_Toc8595020)

# 第1章 引 言

## 1.1 编写目的

本文档适用于软件工程实训平台开发小组成员使用，报告融合了软件工程规范要求的远景、需求分析、总体设计、详细设计及软件测试计划等文档，并附有主要功能的详细设计说明。根据实训平台项目开发需求与实训管理要求，编写项目软件需求规格书。

本文档的目的是确定系统开发的范围；反映出系统的结构，为开发人员、维护人员、需求人员间提供共同的协议而创立基础，作为软件人员进行设计和编码的基础，对软件功能的实现作使命描述；作为实训小组成员的共同文档，为各小组成员完成任务提供依据。该文档将作为项目的问题定义、可行性研究、需求分析、开发阶段所有项目活动的行动基础，并作为项目团队开展和检查项目的工作依据，以及确定系统测试及验收内容。

## 1.1 项目背景

项目的名称：面向互联网文本的语义知识智能分析实验系统-知识驱动对话

项目的导师：刘小明。

项目的用户：百度

项目的任务提出者：刘小明。

项目的主要承担部门：实训小组。

项目建设背景：

最近几年随着互联网的发展, 尤其是社交网络的广泛普及, 人们从互联网上获取到的信息也越来越多。语言信息又是其中最直接的一种,如何从众多的语言信息中发现对我们来说比较重要的信息便显得尤为重要。

人机对话（Human-Computer Dialogue）作为人工智能领域的重要问题，主要包括自然语言理解, 自然语言之间的推理以及一些常识性知识的应用。从20世纪60年代开始, 人们就开始思考如何让计算机像人类一样去理解语言, 随着最近几年随着硬件技术的进步, 尤其是GPU的发展, 使用深度学习的方法进行人机之间的对话也变得越来越容易, 神经网络模型在解决序列到序列的问题的时候也有着天然的优势。比如, 神经网络可以发现非常多的隐藏特征，而这些特征往往是人类无法识别到的，神经网络模型的泛化能力也比传统方法要强的很多，知识驱动对话正是在这种背景下产生的

人机对话是人工智能的一项基本挑战，涉及语言理解、逻辑控制和语言生成等关键技术，受到了学术界和工业界的广泛关注。

总的来说知识驱动对话是一种人机对话任务，机器根据知识信息构建的图谱进行主动聊天，旨在使机器具备模拟人类用语言进行信息传递的能力。

# 第2章 任务概述

## 2.1 功能需求

构建一个面向多种互联网文本（如 新闻、论坛、帖子、微博、微信）格式文本所蕴含的知识驱动系统，机器根据知识信息构建的图谱进行主动聊天，旨在使机器具备模拟人类用语言进行信息传递的能力。前期以基于现有理论技术，完成初步实验结果（Baseline）；后期针对理论模型的改进方案，实现完成改进方案的实验验证。

## 2.2参与者列表

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 技术水平 | 角色 | 工作描述 |
| 赵向阳 | 项目经理 | 项目管理、前期分析、设计、编码 | 分析系统需求、项目计划、项目团队管理、检查进度、项目后期总体负责、数据的处理、对seq2seq建模学习 |
| 陈金新 | 设计、分析师 | 分析、设计、编码 | 环境配置、使用bert对词列表生成词向量、Model的生成、对seq2seq建模学习、文本生成，模型测试 |

表2.2 项目人员表

## 2.2 模块管理

### 2.2.1数据预处理模块管理

1：对语料数据进行下载。

2：对数据进行数据预处理。

3：加入bert词向量预训练

### 2.2.2 Seq2Seq建模模块管理

1：建立Seq2Seq模型，encoder-decoder

2：引入attention，利用多层感知器（Multi-Layer Perception）计算相似度

### 3.2.3 文本生成模块管理

1：利用输入的knowledge，对下一句对话文本进行生成。

# 第3章 需求固定

## 3.1模块需求分析

### 3.1.1数据预处理

1．从电影领域和明星，包括票房、导演、影评等信息提取的训练语料。

2．每个会话包括对话目标，背景知识和对话。

3．将对话目标、背景知识和对话转化为字典，并利用数字映射每个句子。

### 3.1.2 Seq2Seq建模

1．数据预处理后的预料。

2．根据单论对话每次tgt会话进行迭代训练模型。

### 3.1.2 文本生成

1．进行文本生成

## 3.2具体用例描述

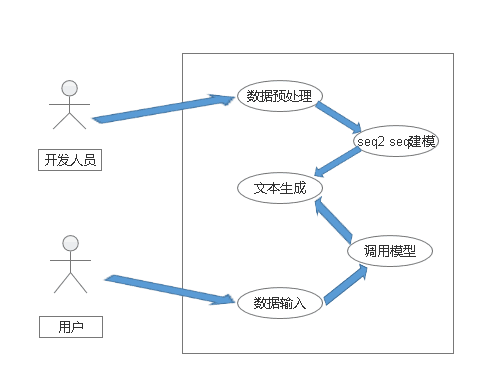


图3.2用例图

## 3.3系统流程分析

### 3.3.1系统主要流程

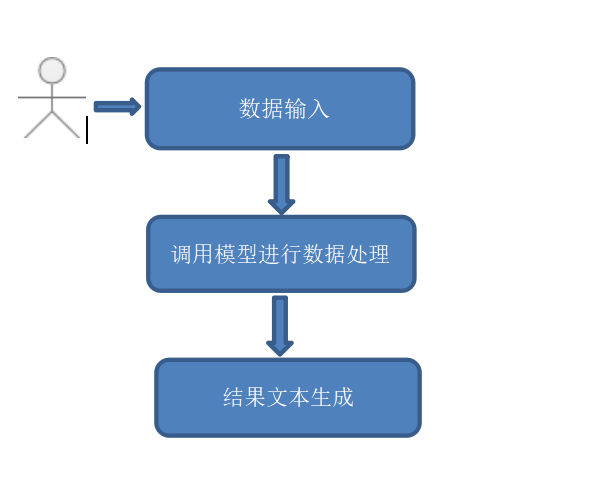


图3.3.1 系统流程图

### 3.3.2系统评测流程图：

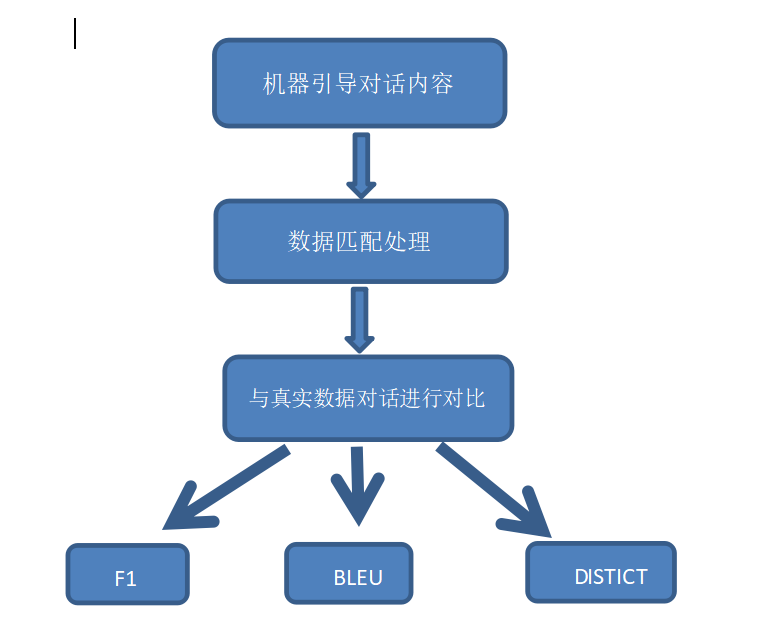


图3.3.2系统评测流程图

## 3.4非功能需求

### 3.4.1访问控制

 只允许在本系统中进行访问。

 非管理人员，无法进入后台修改相关数据。

### 3.4.2信息约束

 使用该系统需要在pycharm中运行，输入训练数据与测试数据，通过模型训练提取出由机器引导的下一段对话内容，通过数据匹配处理，提取出测试数据中应该真实的对话内容，通过经过上下文和知识自然推理出来的下一段对话内容与真实的对话进行比较，可以分别得到F1、BLEU和DISTINCT三个数据的值。进行评估。

该系统分为一个主模块：Seq2seq模型 。

### 3.4.3性能

* 1.精度

 系统的输入：可以识别连贯的文本内容。

 输出数据的精度：

* 2.时间特性要求

选择一个模块进行对文本的分析，软件响应速度达到系统以及相关软硬件环境一般响应速度即可，一般在3s之内，最长不超过10s。数据的更换和传送时间：也在这个范围之内。

由于系统采用B/S结构，因此系统需保证能长期不间断运行。

1. 灵活性

说明对该软件的灵活性的要求，即当需求发生某些变化时，该软件对这些变化的适应能力：

1. 运行环境的变化：该软件适用于现在流行的操作系统。
2. 精度和有效时限的变化：因不同情况而定。

c. 计划的变化和改进：根据用户的需求随时最软件做必要的升级和更新。

# 第4章 运行环境规定

## 4.1软件

|  |  |
| --- | --- |
| **需求名称** | **详细要求** |
| 操作系统 | Windows 7及以上 |
| 操作系统 | Linux |
| 环境管理器 | Anaconda |
| 应用软件 | Pycharm |

## 4.2硬件设备

计算机：最低配置4G内存，双核，10G硬盘。

显卡：最低配置GTX960

显示器：笔记本15.6寸分辨率1280\*800。

服务器：服务器端最低配置16G内存，双核，100G硬盘。

## 4.3网络

WiFi、网线，保证计算机能连接到网络。

网速，1M带宽以上。

## 4.4产品的质量需求

|  |  |
| --- | --- |
| **主要质量属性** | 详细要求 |
| 健壮性 | 任何操作，包括用户操作、系统交互等过程中的错误和异常，只能在相关业务或模块的当前过程中产生影响，并当操作正确时能正常响应。 |
| 可靠性 | 降低有效性的主要因素有：1、系统要经常升级；2、各个系统间消息交互频繁，某个子系统出问题，其它子系统可能就向终端报错。对第一个因素，功能需求中的系统维护可以解决一部分问题；对第二个因素，应保证任何子系统都不会形成单点故障。 |
| 性能，效率 | 1. 系统响应时间：3秒以内； 2. 每秒并发用户数达到200或以上； |
| 易用性 | 因面对的用户可能是对自然语言处理了解不深入的用户，所以对于大多数的交互操作要遵循以下规则：   1. 在界面中对所标注的符号进行解释； 2. 提供部分接口可以让用户进行调用； 3. 界面的语言够简洁易懂； |
| 安全性 | 有关网络环境、操作系统、数据备份、权限管理等方面的安全性要达到高有效性系统的要求，具体要求待定。  应用程序的安全性，主要参考“运维管理”中的“安全管理”。  同时，要满足下面的条件：   1. 密码类重要信息在系统内不能存在明文； 2. 单点刷新过程中的各种口令不能被窃取； 3. 能防止SQL注入、脚本攻击等常见的WEB攻击手段。 |
| 可扩展性 | 可以方便的集成CP，SP系统；  在系统中可以方便扩展新的功能模块； |
| 可移植性 | Server端程序使用python语言开发，并尽量遵从各种python规范，以易于在各种服务器上进行移植；协议方面尽量使用常用的格式。 |