Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2019, 8(3), 102-108
Published Online August 2019 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2019.83013

Hans汉斯

Research on Campus Encyclopedia Knowledge Answering Robot Based on Natural Language Processing

Xiaojin Liu, Dongqing Wu*, Qiaowen Zeng, Qiuyan Yi, Jinwen Kuang, Bowen Chen

College of Computational Science, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong
Email: *rickwu.zhku@qq.com

Received: Jul. 19th, 2019; accepted: Aug. 6th, 2019; published: Aug. 13th, 2019

Abstract

College students often repetitively ask simple questions about campus life. The purpose of this system is to design an intelligent question and answer robot to save time as well as facilitate teachers and students. The continuous development of the information age makes the exploitation and application of computer artificial intelligence more and more extensive. This system uses the technology of word segmentation and short text similarity calculation in natural language processing, and develops a Web application program based on MySQL and Spring Boot. The program is deployed on the Internet and tested in our school with good results.

Keywords

Natural Language Processing, Question Answering System, Campus Knowledge

基于自然语言处理的校园百科知识问答机器人的研究

刘晓瑾，吴东庆*，曾巧文，易秋艳，邝锦文，陈博文

仲恺农业工程学院计算科学学院，广东 广州
Email: *rickwu.zhku@qq.com

收稿日期：2019年7月19日；录用日期：2019年8月6日；发布日期：2019年8月13日

*通讯作者。

文章引用: 刘晓瑾, 吴东庆, 曾巧文, 易秋艳, 邝锦文, 陈博文. 基于自然语言处理的校园百科知识问答机器人的研究[J]. 人工智能与机器人研究, 2019, 8(3): 102-108. DOI: 10.12677/airr.2019.83013

摘　要

大学生在校园里经常会询问一些关于校园生活的简单而又重复的问题。本系统旨在设计一个智能问答系统，达到节省时间，方便在校师生的目的。信息时代的不断发展使计算机人工智能的开发与应用越来越广泛。本系统运用自然语言处理中的分词、短文本相似度计算等技术，基于MySQL和Spring Boot框架开发为Web应用程序，该程序部署到网上并在本校试用效果良好。

关键词

自然语言处理，问答系统，校园知识

1. 引言

近年来，随着信息技术的发展，人工智能成为当今信息时代的研究的一大热点。而智能人机交互中的自动问答机器人是其中最受追捧的研究之一[1] [2]，这些智能机器人可以给人们的生活带来极大的便利。高校的学生，尤其是新生，经常咨询一些关于校园知识的问题，例如某快递的取件点在哪，学校的财务处在哪，某老师教什么课程等等。目前获取信息的方式比较原始和低效，主要有询问同学和老师，查询学校公众号等。如果能够用智能的问答机器人自动回复这些简单重复的信息咨询，不仅可以节省时间，而且也提高了问题的准确度。这大大方便了学生和老师，减轻教务员、辅导员、助班和各部门的工作量。

本系统综合高校学生的需求，可用资源以及团队知识水平等因素，基于关键词提取和短文本相似度计算等技术，开发校园百科知识问答机器人，帮助学生解决一些常见的校园问题。

2. 相关研究与研究基础理论

2.1. 国内外相关研究

本系统实现的自动问答机器人是一个智能的问答系统，能对用户输入的校园问题做出回答。问答系统，也叫做自动问答系统，在其相关领域有较为重要的研究价值。国内外在问答系统这方面的研究都有挺多例子，如 Start、Cortana、Siri 和国内的“度秘”。在线问答方面的研究也有很多，例如文献[4]。

2.2. 分词技术

分词技术就是把一个文本切分成一个个独立、完整、有意义的词组。本系统采用百度的分词技术对文本进行分词、词性标注和专名识别[3]。而常见的几种字符串匹配方法有：正向最大匹配法、逆向最大匹配法、最少切分法。

2.3. 语义匹配模型

本系统所应用的匹配模型 SimNet 是一种有监督的神经网络语义匹配。在语义的表示上 SimNet 依然使用隐式连续向量，基于语义的角度从而获取特征相关的信息，利用分类模型经行分类，后根据分类结

果识别隐式特征[5]。在语义匹配方面，SimNet 应用了深度学习的 End-to-End 模型，该模型特别适用于数据量大的情况。SimNet 主要分为三层，分别是输入层、表示层、匹配层，如图 1 所示。

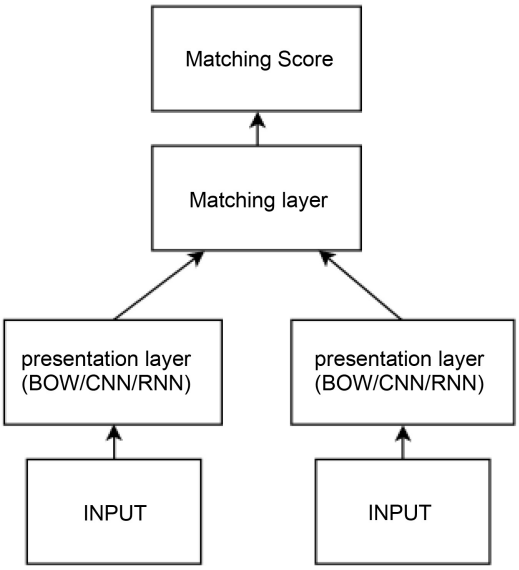

Figure 1. Framework of SimNet
图 1. SimNet 框架图

1) INPUT layer 通过 look up table 将文本词序列转换为 word embedding 序列。

2) Presentation 的功能是构建句子，将独立词语的 embedding 表示组建成为为具有全局信息的一个或多个低维紧凑的语义向量。

3) Matching layer 的功能是匹配。利用表示层生成的文本向量进行相似度计算，这里共有两种匹配算法，分别是 Representation-based Match 和 Interaction-based Match。

① Representation-based Match

该方式重点是对于表示层的构建。有了向量后就可以进行匹配计算。常用的匹配计算有余弦相似度和多层感知网络(MLP)，其中余弦函数的使用更频繁。余弦相似度侧重于在方向上区分向量间的差异，对数值不敏感，这个特点适合用于文本内容的区分。而 MLP 是通过数据训练拟合出一个得分，该得分基于匹配度，匹配度越高，分数越高，这种方式相对而言拟合能力强，灵活度高，相对与 Cosine 要求更高，处理也更复杂。二者对比如图 2 所示。

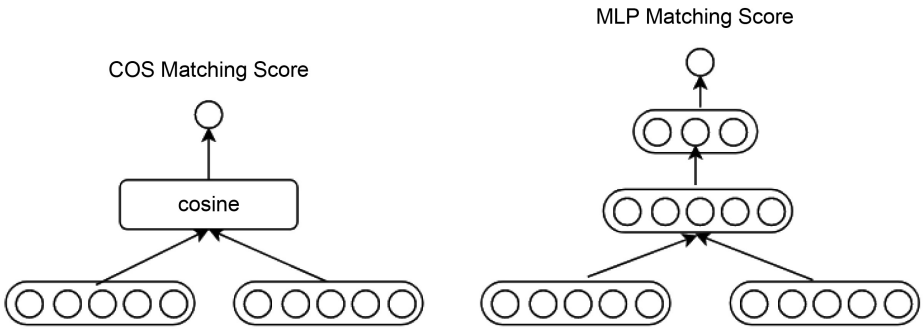

Figure 2. Cosine (left) vs MLP (right)
图 2. Cosine (左)与 MLP (右)比较

余弦相似度公式：

$$\cos\theta = \frac{X_1X_2 + Y_1Y_2}{\sqrt{X_1^2 + Y_1^2} \times \sqrt{X_2^2 + Y_2^2}} \tag{1}$$

对 n 维向量依然成立：

$$\cos\theta = \frac{\sum_{i=1}^{n}\left(A_1 \times B_2\right)}{\sqrt{\sum_{i=1}^{n}A_i^2} \times \sqrt{\sum_{i=1}^{n}B_i^2}} = \frac{A \cdot B}{|A| \times |B|} \tag{2}$$

MLP (Multilayer Perceptron)多层感知器，是一种前向结构的人工神经网络，映照一组输入向量到一组输出向量，通常使用反向传播算法来训练 MLP。

② Interaction-based Match

该匹配建模更加精准效果更好，相对的计算成本更高，一般应用于对匹配精度有要求的场景，实际使用的比较少。

3. 系统设计(Systematic Design)

本问答系统命名为“仲园百晓通”，提供有普通用户端和管理端，分别对应两类用户：仲恺在校生、系统管理员。对于不同类型的用户有不同的功能使用权限。整个系统的系统框架图如图 3 所示。

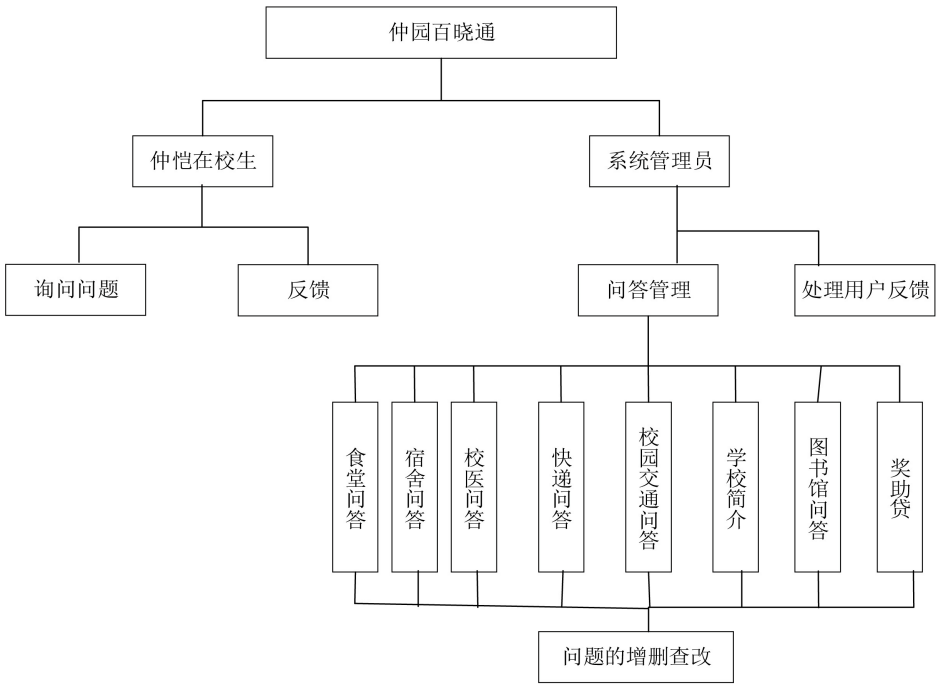

Figure 3. System function frame diagram
图 3. 系统功能框架图

4. 系统实现

4.1. 开发工具及运行环境

本系统采用 Spring Boot 框架为后端开发技术，前端用 CSS 和 DIV 的方式设计网页，数据库采用了 MySQL。整个系统基于 B/S 架构。

4.2. 核心部分实现

1) 检索流程实现

前台用户输入问题之后，程序首先对所输入的问题进行分词，若是所问问题包含数据库提供的关键词，将获得问题所在的表名。否则，系统将提示错误。

2) 分词与问题匹配实现

本系统调用了百度自然语言处理的“分词”和“短文本相似度”两个接口。程序根据用户输入的问题进行分词，判断得到的结果是否含有关键词，如有关键词，则得到关键词所在的目标查询表。

4.3. 功能实现

系统操作流程图如图 4 所示。

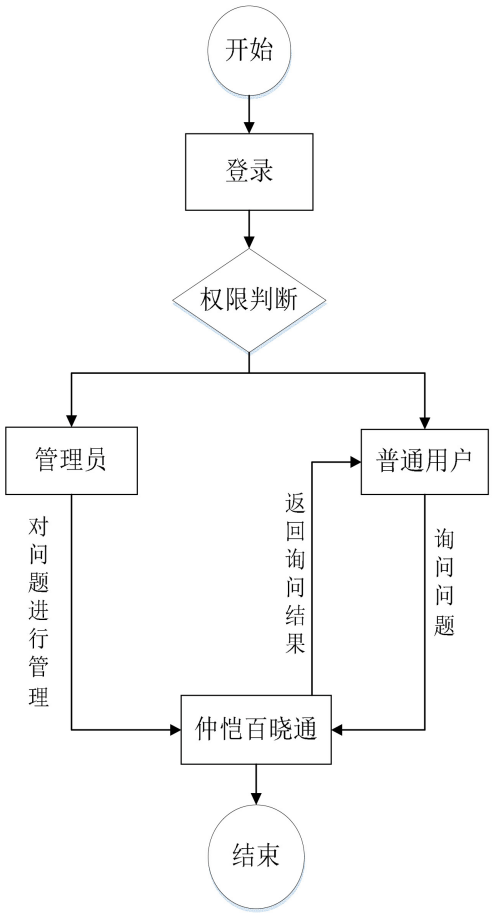

Figure 4. System operation flow chart
图 4. 系统操作流程图

用户在浏览器中输入正确的 URL 地址后，就可以访问系统首页，如图 5 所示。

输入账号密码，以普通用户的身份登录。进入用户主页后可以对系统进行问题询问，获得相应的答案。如图 6 所示。

输入账号密码，以系统管理员的身份登录。进入管理员主页后可以对数据库中不同类型的问题对进行管理。以快递信息管理模块为例，如图 7 所示。

Figure 5. The system's front page
图 5. 系统首页

Figure 6. Quiz interface
图 6. 问答界面

Figure 7. Express information management module interface
图 7. 快递信息管理模块界面

5. 结论

本系统旨在为全校师生提供一个官方的信息来源，在保证信息正确性的同时，达到了节省询问时间，简约询问方式，方便在校师生的目的，对提升校园百科信息服务水平有显著效果。目前本系统已经在本校推广试用，获得好评。

基金项目

2018 年国家级、省级大学生创新创业训练计划项目(201811347024、201811347086)；2019 仲恺农业工程学院校级质量工程项目(KA190573919)；2019 广州市哲学社会科学发展“十三五”规划 2019 年度课题(2019gzgj125)。

参考文献

[1] Tse, R. and Campbell, M. (2018) Human-Robot Communications of Probabilistic Beliefs via a Dirichlet Process Mixture of Statements. *IEEE Transactions on Robotics*, **34**, 1280-1298.

[2] Sun, S., Chen, L. and Chen, J. (2017) A Review of Natural Language Processing Techniques for Opinion Mining Systems. *Information Fusion*, **36**, 10-25. https://doi.org/10.1016/j.inffus.2016.10.004

[3] Singh, B. and Singh, U. (2017) A Forensic Insight into Windows 10 Cortana Search. *Computers & Security*, **66**, 142-154. https://doi.org/10.1016/j.cose.2017.01.007

[4] 张馨雨. 群聊话题检测技术研究[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2016.

[5] 吴育良. 百度中文分词技术浅析[J]. 河南图书馆学刊, 2008, 28(4): 115-117.