基于自然语言处理的聊天机器人：人大附中 2020-2021 学年度研究性学习机器学习领域开题报告

安阳	彭健坤	王家瑞
人大附中	人大附中	人大附中
18210230897	18801291386	15201011545
shuijingty@outlook.com	pjk2013@163.com	XGHDSGSDH@163.com

1 引言

1.1 选择本课题的背景及目的

在现在这个快速发展的时代中，人们对于传统社交的需求有所降低，宅在家似乎成为了很多人的生活方式。尤其在大城市，青年人的独居率逐年上升，他们都倾向于避免很多无意义的传统社交方式，而集中于自己的小世界，更何况在新冠疫情期间，工作、学习都在家里已经成为常态。而且近年人工智能语言处理的快速发展已经使与人工智能进行阳间对话成为可能。与他人沟通是人类最原始的需要，所以我们想做一个基于自然语言处理的聊天人工智能，在需要的时候缓解孤独，也许还可以在一些特殊疾病的患者身上得到更多的应用。

1.2 选题意义

- 可以进行情感分析并进行对应的回复，可以与人进行基本的对话，在残疾人帮扶，辅助老年人生活等方面大有作为。
- 与其他智能设备结合，构建更加智能的智能家居/智能设备系统。

2 文献综述

2.1 背景介绍

自然语言处理（NLP）是计算机科学的一个分支，分为自然语言理解（NLU）和自然语言生成（NLG）两个方面。其目的在于帮助机器理解、处理和分析人类语言。其核心算法在于对文本的分析与处理，分析文本上下文相关性的逻辑，进而提取出其中的特点，辅助人们处理大规模的文本信息。在聊天机器人方面，本文将专注于 NLP 和关于机器问答的 NLG 实现部分。

NLP 的实现技术发展大致经历三个历程：规则主导的语言处理（-1990）、基于统计的语言处理（1990-2014）、深度学习实现语言分析（2014-）[8, 11, 13, 14]。随着计算机的发展及通信智能设备的普及，计算能力、可收集数据量的增加，自然语言处理的开发愈加依赖于数据驱动的方法，这些方法有利于构建出更加健壮、鲁棒性更强的模型，且

相对于传统方法，利用大数据训练的深度学习的新方法准确率更高，且更加易于设计。当下，利用与深度学习与神经网络结合的模型实现自然语言处理功能已经成为最为主流与高效的处理方式。

2.2 数据表示

为了更好地提取出语段中诸如用词习惯、语段倾向等特征，需要用一或几种表示方式来表示出语段的意义特征。用来数据表示的方法数不胜数，但大都是围绕词汇进行，在此摘取若干进行分析。

2.2.1 One-hot Representation

在基于规则和统计的自然语言分析和处理方法中，最为经典的方法便是 One-hot Representation 表示方法。在此方法中，每个词汇被表示成一个由“0”和“1”组成的字符串或向量，如 [00100000$\cdots$] 。其中，代表本词汇的一位被赋为 1，其它位被赋为 0。这样，每一个句子、每一篇文章都可以被表示为一个向量的有序集合，这便形成了一个从自然问题到数学问题的转化。而在其基础上进行诸如支持向量机（SVM）等操作也更加方便。但是其缺陷也是显然的：其向量方式只能表示单独的词汇，并不能够体现词汇之间的联系，无论是同义词还是反义词，词的几乎一切特征都被忽略不计。

2.2.2 词向量模型

在 One-hot representation 之后，Hinton 提出了一个在其基础之上表示词的方式，即词向量模型 [2]。在此模型中，每个词汇被表示成一种高维的实数向量，如 [0.274, 0.238, –0.064, 0.919, $\cdots$] 。在经过训练后，在用此方式表示的词汇向量集中，形如“皮鞋”和“皮靴”的词汇的空间距离将会远小于“皮鞋”和“水牛”的距离。这样，我们就可以计算向量之间的几何距离或余弦距离来评估词汇间的相似程度。在当下，词向量模型被广泛地运用在大量的 NLP 实现中，是最为常用和经典的数据表示方法。

2.2.3 词袋模型

在自然语言处理领域，词袋模型一般用来处理表示文档。在词袋模型中，词袋模型用文档中所有词汇的集合及每个词汇的出现次数来表示整个文档。如 [”*Talk*” : 1, ”*Computer*” : 2, $\cdots$] 。在构建字典完成之后，便可以把每个词映射到对应的编号上。然后成为两个向量之间的映射关系。这样把每个文档转换为两个向量后，便可以方便的利用计算机进行文档分类等任务。

2.3 NLU 的一般实现方法

2.3.1 分词

分词部分的主要任务是将连续的字序列分割为词语序列，以便于语义分析。目前主流的分词方法是 Sun Junyi 开发的 jieba 库，他的团队利用字典树 (trie 树) 来找到所有的可能词语，在子序列构造 DAG，并通过预设的字典词频依靠动态规划算法实现概率最大路径的寻找。在涉及到预设词典未收录的词时，主要用到了 HMM 算法的预测问题模型，并使用 Viterbi 算法来求解最优预测结果。利用 HMM 模型进行分词，主要是将分词问题视为一个序列标注（sequence labeling）问题，其中，句子为观测序列，分词结果为状态序列。首先通过语料训练出 HMM 相关的模型，然后利用 Viterbi 算法进行求解，最终得到最优的状态序列，然后再根据状态序列，输出分词结果。基于以上算法即可高效且准确的分词 [12]。

2.3.2 词性标注

词性标注即对句子中每个词的词性进行确定，利用形容词、动词、名词等标签对词汇进行标记。关于中文的词性标注，复旦大学郑晓庆博士等人在 2013 年给出了一种利用神经网络框架进行中文的分词及词性标注的方法 [10]。相比于原中文依靠词典标注的传统方法，郑晓庆博士等人提出的新方法极大降低了时间复杂度和实现难度，同时保持了较优秀的效果。同时，他们提出了中文字向量的概念，使得利用大量未标注的中文数据进行神经网络训练成为可能。

2.3.3 句法分析

句法分析的主要任务是将句子依照句法分割为不同的句子结构，如主谓宾及各种从句等，并确定

句子中各个组份的逻辑关系，为接下来更加复杂的处理做准备。一般分析完成后会以句法树的形式给出句法分析的结果。

目前的句法分析方法不一，各有特点。其中斯坦福大学的 Richard Socher 等提出了一种组合向量语法模型 [5]（Compositional Vector Grammar, CVG）用于预测句法结构，并结合了上下文概率语法结构（Probabilistic Context Free Grammars, PCFG）与递归神经网络结合，获得了部分性能上的提升，其中用于评估模型的 $F1$ 值达到了 90.4%，并且提高了约 20% 的训练速度，是一种相对较为优秀的模型。同时，他们分析了当前存在的句法分析工具 [7]，并在 2012 年再次优化循环神经网络 [6]，使得算法准确率进一步提升。

2.3.4 词义学习

词义学习涉及利用文本内容对词向量进行调整及优化，即对词义的学习。但是与文本中依托上下文决定的词义不同，目前的词向量仍然较为独立地存在，难以联系上下文，且存在一词多义难以表达的问题。Huang 等 [3] 在 Collobert 等 [1] 的基础上利用深度神经网络同时综合全局和本地信息，能够结合上下文更好地理解词义，同时改进了词向量模型，使得一个词能够对应多个词向量，从而解决了对一词多义的支持，包含更加丰富的语义信息。在实验中，Huang 等的方法与人工标注语义更加相近。

2.4 机器问答的实现理论

机器问答任务对于 NLP 而言极其困难，且此领域相对而言较为年轻，只有约 10 年左右的历史，但这与其未来巨大的潜力并不相矛盾，在近十年间，NLP 技术发展迅速。2014 年，日本庆应义塾大学的相良司和萩原将文提出了机器问答任务的神经网络实现的一般流程 [4]，而 2015 年 Facebook 更是引入了记忆神经网络（Memory Networks）[9]，在语句经过语义分析和筛选之后，先验事实被输入到神经网络中，并在大量的问答测试中表现良好。

2.5 目前成果

目前市面上自然语言的研究成果众多，但主要集中在英语方面。在聊天领域，以微软公司推出的人工智能“小冰”和“小娜”最为知名，以及较为知名的 Replika 等。其较高的对话智能、遣词造句甚至作诗的能力均让人叹为观止。

另外，人工智能在撰写新闻稿件方面也已有很大成果。美联社的 WordSmith、腾讯的 Dreamwriter、纽约时报的 Blossom 等均是目前存在的很优秀的撰稿机器人，其撰写的速度、客观性等均优于人类撰稿人。

2.6 对于目前技术的评价与展望

自然语言处理技术的应用前景在可见的未来是非常广泛的。其可以完成文本分类、信息抽取、情感分析、机器翻译、问题回答、文件摘要、语音对话等诸多任务，优秀的 NLP 语言分析系统在未来必然大有可为。

需要指出的是，关于语言处理的算法多基于英语的单一语言，除机器翻译之外的方法都对其他语言的支持较为薄弱。而且，目前的自然语言处理对于文章的情感分析仅流于表面的情感的二元判断，而不能进行与原句结合的，更加有机深入的理解与分析。对于文章生成方面，其在客观表述方面有较好的表现，但是在情感表达方面仍有较强的拓展空间。

3 研究目标与内容

3.1 研究目标

基于以上算法，实现一个可在电脑端运行的聊天人工智能。

3.2 研究内容

对上述算法进行研究和尝试，实现以上算法的有机结合。

4 时间安排

9 月-10 月	查阅文献，明确研究问题和研究方案、工具
11 月	开题——撰写文献综述和开题报告，汇报开题报告
12 月 1 日—寒假前	预调查、分析数据、改进问卷
开学—3 月 1 日	数据处理完成
3 月 1 日-4 月 20 日	撰写报告并修改

5 致谢

本组成员们衷心感谢中国人民大学附属中学计算机与科学技术学科梁霄老师的热情帮助与提供的大力支持。

参考文献

[1] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery.

[2] Hinton. G. E. Learning distributed representations of concepts. In Eighth Conference of the Cognitive Science Society, 1989.

[3] Huang Eric, Socher Richard, Manning Chris, and Ng. Andrew. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, pages 873–882, 2012.

[4] Tsukasa Sagara and Masafumi Hagiwara. Natural language neural network and its application to question-answering system. Neurocomput., 142:201–208, October 2014.

[5] Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. Computer Science Department, 2008.

[6] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. EMNLP-CoNLL '12, page 1201–1211, USA, 2012. Association for Computational Linguistics.

[7] Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks, 2010.

[8] Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. Natural language processing advancements by deep learning: A survey, 2020.

[9] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks, 2014.

[10] Zheng Xiaoqing, Chen Hanyang, and Xu Tianyu. Deep learning for chinese word segmentation and pos tagging. In Conference on Empirical Methods in Natural Language Processing, pages 647–657, 2013.

[11] 余凯, 贾磊, 陈雨强, and 徐伟. 深度学习的昨天、今天和明天. 计算机研究与发展, 050(009):1799–1804, 2013.

[12] 刘毛毛. jieba 分词的原理, 2020.

[13] 奚雪峰 and 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, (42):1445–1465, 2016.

[14] 林奕欧, 雷航, 李晓瑜, and 吴佳. 自然语言处理中的深度学习：方法及应用. 电子科技大学学报, 46(6), 2017.