硕士学位论文

交互式问答中的语句关系识别方法

METHODS OF SENTENCE RELATION RECOGNITION FOR INTERACTIVE QUESTION ANSWERING

陈静

哈尔滨工业大学

2017 年 6 月

国内图书分类号：TP391.3　　学校代码：10213
国际图书分类号：004.9　　密级：公开

工学硕士学位论文

交互式问答中的语句关系识别方法

硕士研究生：陈静
导　　师：陈清财 教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：深圳研究生院
答辩日期：2017 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.3

U.D.C: 004.9

Dissertation for the Master Degree in Engineering

METHODS OF SENTENCE RELATION RECOGNITION FOR INTERACTIVE QUESTION ANSWERING

Candidate: Jing Chen

Supervisor: Prof. Qingcai Chen

Academic Degree Applied for: Master of Engineering

Speciality: Computer Science and Technology

Affiliation: Shenzhen Graduate School

Date of Defence: June, 2017

Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

随着互联网技术的发展和信息量的迅速增长，人们迫切需要一种准确、高效的信息获取方式。从搜索引擎到智能交互式问答系统，信息的获取方式越来越接近于自然交互。一方面因为海量数据的出现，另一方面因为机器学习和自然语言处理等技术的长足进步，问答系统进入了面向各领域、基于自由文本和异构信息、基于生成式的智能交互式问答发展阶段。与搜索引擎不同的是，用户无需在多条候选文档中选择，问答系统能更好的理解以自然语言形式描述的问题，同时返回简洁精确的答案。随着 Siri 和 Watson 的成功问世，智能交互式问答系统成为了近年来的一个研究热点，在商业领域也越来越具有代替人工客服的潜力。然而，要构建更加智能的交互式问答系统，从已有的客服日志中学习知识就显得非常重要，而如何从复杂的交互式问答客服日志中识别问句与答句之间的匹配关系以及连续语句之间的补充关系则成为了构建学习系统的关键。本文主要针对交互式问答中的语句匹配关系识别和补充关系识别进行了研究。

针对客户问句与客服回答之间的匹配问题，本文分别构建了基于 CNN 的语义匹配模型和基于 RNN 的生成模型，模型的输入层是句子的词向量矩阵，输出层是问答匹配的置信度。分别在 Semeval-2016 社区问答数据和在线客服对话数据上，进行了不同模型的性能对比。同时对问句的完整性、生成模型的不同结构、阈值选择以及客服数据的抽取方式等进行了对比实验分析。实验结果表明，在社区问答数据中，本文中基于 CNN 的匹配模型优于 RNN 生成模型；在客服对话数据中，基于 RNN 的序列学习模型能够更好的学习到场景对话中的上下文信息。在基于每轮对话且问句完整的数据上，MAP 达到了 84.41%。

针对交互式问答中连续语句之间存在的上下文相关联的潜在语义补充关系，本文研究了句子补充关系的识别。在深度模型上，构建了并行 CNN 和串联 LSTM 对句子对进行抽象语义特征提取和建模。分别采用支持向量机、基于 CNN 的模型和基于 RNN 的模型，对句子对的补充关系进行分类。实验结果表明，基于 CNN 的识别方法优于其他对比方法，其 F1 值达到了 67.8%。最终，将补充关系识别和匹配关系识别相结合应用于交互式问答语义匹配。

关键词：问答匹配关系；补充关系；卷积神经网络；循环神经网络

Abstract

With the development of Internet technology and the rapid growth of information scale, people urgently need an accurate and efficient way to obtain information. From search engines to intelligent interactive question answering systems, the ways of searching information become more and more close to the natural interaction. Because of the emergence of massive Internet data and the progress of machine learning methods and natural language processing technology, the QA system for free text and heterogeneous information occurs in some fields, and it aims to be an intelligent system based on generative models. Different from search engines, the question answering system can understand the problem described in the form of natural languages better, and the system returns concise and accurate answers instead of some related documents. With the occurrences of Siri and Watson, the intelligent interactive question answering system has become a hot research topic in recent years. In the business field, it is also more and more potential to replace manual customer service. However, to construct an intelligent interactive QA system, it is very important to learn knowledge from the real customer service logs. And how to recognize the matching relations between questions and answers and the completion relations between continuous sentences from those complex interactive logs has become the key to construct the learning system. This paper mainly studies the question answer matching relation recognition and the completion relation recognition in the interactive question answering.

For question answer matching, this paper constructs two kinds of models including the semantic matching model based on CNN and the generative model based on RNN. The word vector matrices of sentences are feed into the input layer of models, and the output are the question answer matching confidence. The comparative experiments of different deep learning models are carried out based on the Semeval-2016 community question answering data and the online customer service dialogue data. Furthermore, this paper analyses the experiment results based on the completeness of the questions, different structures of the generative model, the threshold selection and the methods of data extraction. It was found that on the community question answering data, the CNN model performs better than the RNN model. On the customer service dialogue data, the RNN based sequence to sequence model can learn scene information from the dialogues better. In the one round based data with a complete question, MAP reaches 84.41%.

In view of the potential semantic completion relations between continuous sentences in the interactive questions answering, this paper studies the recognition of

sentence completion relations. The paper constructs the parallel convolutional neural networks and series LSTM to extract high-level semantic features and model the sentence pairs. And this paper uses the Support Vector Machine, the CNN based model and the RNN based model to recognize the completion relations between sentences respectively. Experimental results show that the CNN based method outperforms other comparative methods and achieves the best F1 value of 67.8%. Finally, the question completion relation recognition and the matching relation recognition are combined to be applied to the interactive question answer semantic matching.

Keywords: question answer matching relations, completion relations, convolutional neural network, recurrent neural network

目 录

第 1 章 绪　论

1.1 课题来源

本课题来源于哈尔滨工业大学深圳研究生院智能计算研究中心的合作项目，主要目的是在基于规则的客服问答系统上，通过深度学习的方法，进行问答语句关系的识别，以帮助抽取人工对话的知识对，改善系统的问答交互效果。

1.2 课题研究的背景及意义

随着互联网技术的迅速发展和数据量的剧增，人们迫切的需要一种高效、准确的方式从海量的信息中寻找到自己想要的答案[1]。搜索引擎是我们常用的一种信息获取手段，但是它需要用户在其返回的一堆相关文档中进行挑选，并不能返回给用户精确的答案，无法满足用户对信息获取快速、便捷、准确的要求[2]。从黄页查询到搜索引擎，到专家系统，再到智能交互式问答系统，不难看出，人们要求信息获取的方式越来越类人化，无需思考使用什么样的关键词组合，无需在大量的信息纯度不高的候选答案中仔细挑拣，仅使用自然语言的提问方式，就能获得理想的答案。由于这些原因，无论是在学术研究上还是在工业应用上，自动问答已然成为了一个新的研究方向和热点。

智能交互式问答系统将成为新一代的信息检索和自动问答的基本形态。随着传统行业与互联网行业的深入结合，大量新的产业业态出现，尤其是电子商务得到了空前的发展，随之带来的是对人工客服的极大需求。但使用人工客服需要较高的劳力成本和培训成本，同时并不能像机器一样为客户提供全天候的高效率实时服务，因此智能客服问答系统已被逐渐应用于商业化领域。目前，部分客服已减少或取消人工电话客服，而改用机器和人工结合的文本聊天方式，或者在客户与人工客服交互前，让他们对咨询内容做出选择，同时咨询结束后，要求客户对交互过程给出评分。其目的就是希望通过这种方式搜集更多的真实场景中用户与人工客服的自然交互数据，为智能客服的发展提供充足的研究数据，以期未来能够达到客服的机器化、人性化、及时性以及低劳力成本。不难看出，智能交互式问答系统具有极大的代替人工客服的趋势，这也是我们课题研究的意义所在。

然而，目前市场上的智能客服系统，并没有完全达到我们所希望的智能化，更多的，选择基于规则、基于人工知识库的方式去实现，这不仅需要大量的人力和物力去完成，同时在当今这个信息更新速度极快的时代，自然语言的多样

性特点导致这样的工作无穷无尽。如何从大量的基于自然语言文本的人工客服对话日志中学习到有意义的问答对是问题的关键。随着数据规模和技术的发展，我们期望通过一些能够自动学习语言特征和对话场景信息特征的方法，帮助我们完成问答对的自动抽取。在此过程中，对话数据中语句间存在的两种重要关系——匹配关系和补充关系的识别成为了问题的重点。深度学习是一种典型的能够自动学习特征的方法，因此，本课题在人工客服对话的基础上，构建深度学习的相关模型对人工对话中的问答匹配关系识别进行研究，扩展了生成模型在语义信息上的自动归纳的应用，同时由于对话中存在单个语句遗漏信息的现象，对自然对话中语句补充关系的识别进行了研究。

1.3 国内外相关技术研究现状分析

1.3.1 问答系统与评测现状

在 20 世纪初，计算机还没有出现的时候，图灵（A. Turing）就提出了著名的“图灵实验”来检验计算机智能的高低。Turing 实验的提出曾经让专家系统在自然语言处理领域的研究非常流行，这也就是早期的问答系统，由于早期技术条件有限，主要局限于特殊领域。比较有名的专家系统，比如上世纪 70 年代研发的 LUNAR[3]，用于回答阿波罗的月球岩石样本的地质分析问题。随着大数据处理技术和文本处理技术的兴起，这种传统的问答系统的技术研究逐渐沉寂。由此，传统意义上的人工智能问答系统已经不再流行，随着大规模真实语料库的出现，兴起了基于大规模语料的机器学习和统计研究。在此期间，信息抽取成为了一个比较热门的方向。与之相关的竞赛也逐渐兴起，比较有名的比如受美国 DARPA 项目资助的信息理解会议(Message Understanding Conference, MUC)，与以往的自然语言处理研究不同的是，其更倾向于实用化，一下缩短了语言处理技术与实用的距离。这一先行的技术研究，为新一代的问答技术的起步起到了很重要的引领作用。1993 年第一个面向互联网的基于自然语言的问答系统 START 诞生，1995 年第一个聊天机器人 ALICE 诞生，2011 年 IBM 公司研制的 Waston，在美国知名智力比赛节目中打败了冠军选手，2014 年一台超级计算机号称通过了图灵测试，在 5 分钟的沟通中，让 33%的测试者相信了它是一名 13 岁的男孩。目前比较著名的问答系统包括谷歌公司的 Google Now、苹果公司的 Siri、百度的度秘、微软的小冰等。

经过了几十年的发展，问答系统越来越类人化，趋于人类的交互模式，慢慢向图灵测试的目标靠近。自动问答系统从专家系统、聊天机器人发展到了各行业服务型的智能交互式机器人。问答系统的类型也变得越来越多，从技术角

度来说，包括基于规则的问答系统、基于知识库的问答系统[4]、基于序列到序列的自由文本问答系统等。近年来，无论是在学术界还是在工业界，交互式问答的相关技术都受到了广泛关注并渐渐得到应用。正如 TREC-8 的广告语：“用户有问题，他们需要答案”所说，非常直接恰当的描述了问答系统的价值定位和终极使命。

随着自然语言处理技术的发展和大量数据的涌现，问答系统研究已经逐渐从原先的信息检索方向中独立，成为了新的研究方向，与之相关的评测也慢慢涌现。从 1999 年的国际文本信息检索会议(Text Retrieval Conference, TREC)开始，已经专门加入了问答评测这一任务，在其推动下，自动问答技术得到了很大的发展[5]。与 TREC 同是处理欧洲语系的评测还有 CLEF 中的 QA 分支[6]。由于他们处理的均为欧洲语系，在技术上不一定适用于亚洲语系，比如中文、日文等。在 2003 年，日本国家科学咨询系统中心 NTCIR（NACSIS Test Collections for IR）首次举办了日文的问答评测，随后，于 2005 年增加了中文问答评测[7]。目前而言，解决效果最好的一类是事实类问题[8]，取得了不错的效果，其他类别的问题相对于事实类问题的研究进展较缓慢[9]。2015 年，TREC 的 QA Track 开展了一个针对于 LiveQA 的任务。以往 TREC 处理的主要是虚构的事实类问题，与之不同的是，15 年的任务，他们首次使用了来自 Yahoo 上提交的未被回答的真实用户所提的问题。这更加拉近了问答任务评测及相关研究与实用的距离。同样，在 SemEval-2015、Semeval-2016 任务三中，给出了问句匹配，答案选择和扩展的问句匹配和答案选择评测。不难看出，问答相关技术的研究是当今的一个热门方向，同时也在积极的向应用方向靠近。

纵观自动问答的发展历史，其中部分问答系统有如下几个：

（1）START　世界上第一个基于互联网的问答系统，自 1993 年开发以来一直运行至今，由麻省理工大学计算科学与人工智能实验室研发。与信息检索不同，START 的目标是只返回唯一的答案，而不是一组相关链接。目前为止，该系统能够回答数百万的问题包括地理、电影、人文、科技、历史、词典定义等。

（2）AnswerBus[10]　是一个基于互联网的开放域问答系统，自 2001 年开始运行至今。其实现了句子层面的信息检索，接受用户输入的六种语言的问题，通过从 5 个常用搜索引擎中选择 2-3 个合适的搜索引擎并将自然语言问题转化为合适的检索式进行检索，返回用户满意的相关答案。使用了动态命名实体抽取、QA 词典、浅层 NLP 分析等技术。当时在 200 个问题测试中平均响应时间为 7 秒，检准率达到了 70.5%。

（3）MURAX[11]　是一个基于封闭类问题的问答系统，使用在线的百科全

书作为知识库来回答一般性的知识问题。给定一个问题，将问题表示为带距离约束的布尔查询，利用一个浅层的句法分析器按照百科全书章节中的答案与问题中的名词短语共现度和答案类型来抽取潜在答案。

（4）Webclopedia[12]　该系统使用一个问答类别分类器进行覆盖，使用一个句法语义解析器对问题进行解析并获得查询获得相关文档，然后对相关问答进行分割，集合词和解析树的信息对答案信息进行判断和抽取。主要使用了词级别的信息检索技术和句法语义级别的自然语言处理技术。

（5）IBM Watson　于 2011 年发布，是认知计算的杰出代表。能够利用自然语言处理技术、信息检索技术、机器知识表达、推理和学习技术，对数以百万计的字典、百科全书、戏剧等非结构化信息进行知识构建，使用多种算法投票选择最佳答案。它具有较强的理解、推理和学习的能力，为用户提供交互式的体验。

（6）微软小冰　于 2014 年发布，是一款智能对话机器人。其主要是搜集大量的网络数据，使用自然语言处理、大数据、机器学习和深度模型相关技术，将资料转化为问答语料库，学习语境和语义以达到人机自然交互的效果。同时，还会像人一样，每天学习更新知识。据报道，微软小冰于 2015 年以见习主播登陆东方卫视主持每日天气播报。

（7）Siri　是苹果公司的一款智能语音助手，于 2011 年在 iphone4S 产品上发布。Siri 可以支持自然语言输入，支持多种语言，提供搜寻餐厅、订票、闲聊等服务，给出交互式的应答。

相对于国外而言，国内在问答系统方面的研究仍有一定的差距，主要是由于中文句法语法和表达比较复杂，在很多方面，与欧语系有所不同，中文的信息处理有其自己的特点，存在一定的技术难度，国外一些相关的成熟技术和研究成果不能被直接利用。同时目前关于中文的评测会议比较少，而 TREC、Semeval 等国际会议，就一直都有针对英文问答系统的评测语料和评测标准。不过，近年来，国内对问答技术的研究发展迅速，清华大学和北京大学、哈尔滨工业大学、中科院计算所等单位先后参加了 TREC/QA，Semeval 的 QA 评测，都取得了不错的成绩，同时各大高校、公司也在努力推进中文问答相关的评测、比赛和应用。

学术界来说，比较有名的问答系统有：哈工大信息检索实验室的开放式问答系统，中科院计算技术研究所的 NKI 的问答系统，北京大学计算语言研究所和香港科技大学的提问式搜索引擎 Weniwen，台湾国防大学管理学院研制的中文问答系统 CQAS 等。NKI 主要接受用户以自然语言的方式进行地理、人物、中医等相关知识的提问；CQAS 主要使用命名实体的关系串列方式进行答案的

搜寻。

商业领域来说，自动问答系统也是应用得越来越广泛，比如小 i 机器人、京东 JIMI，360 问答，度秘。

（1）**JIMI**　　JIMI 是京东自主研发的用于京东商城的交互式客服机器人，于 2014 年上线，主要处理购物各个环节的客户服务，同时也包括闲聊等话题。JIMI 会对用户问题进行命名实体抽取、意图识别和分类，然后在相应的类别中确定候选答案并进行选择。使用了深度神经网络、用户画像以及一系列自然语言处理技术对意图识别和答案选择两大问题进行解决。目前，京东更在积极推动用户使用和反馈以搜集更多的用户数据进行模型的学习。在很大程度上，JIMI 缓解了京东人工客服的压力。

（2）**度秘**　　于 2015 年在百度世界大会上推出，以对话式智能秘书为定位。能够进行多模态的多轮对话交互，具有设备控制、日程管理、信息查询、生活服务相关方面的能力，结合了语音识别、图文识别、自然语言处理、机器学习、深度网络等多种技术，在百度强大的用户数据上不断学习，为用户提供满意的答案。

与传统的信息检索和问答系统相比，现在的问答系统越来越趋近于人与机器的自然交互，同时交互方式也不仅局限于自由文本，还包括语音、图文等，无需用户从一堆候选答案中进行挑拣而是返回给用户唯一的答案，技术上的发展也更加丰富和成熟，比如目前比较热门的深度模型，得到了进一步的应用。目前，各大科研院所在问答方向上的研究主要包括问句理解、问答匹配、问句补充、用户意图分析等方面，重点和热点已经转移到了多轮对话的相关研究上，深度学习方法得到了广泛的研究和使用。

1.3.2 传统自动问答技术研究现状

自动问答的相关任务主要包括问句理解，答案选择，问句匹配、问答匹配等方面。从技术上说，目前的自动问答技术主要是在基于检索、基于知识库、基于 seq2seq 等方面来进行的。

2006 年，Feng 等人判别用户问题的兴趣，从已标注的 1236 个归档的学生在线讨论和 279 个课程文档中挖掘合适的答案，生成类人的回复。其利用信息检索和自然语言处理技术，比如有余弦相似度计算，完成对在线讨论的分析、问句的匹配和答案抽取的工作，模仿人进行在线交互学习[13]。2007 年，Huang 等人提出了一种级联框架完成问答对的抽取工作。从在线论坛中提取高质量问答对作为聊天知识库以高效的支持聊天机器人的构建。首先基于结构和内容的相关性使用 SVM 分类器进行分类，再基于内容质量使用 SVM 进行排序选择排

名靠前的问答对。论文中以电影论坛为例，进行了问答对的抽取工作，取得了很大的效果提升[14]。2008 年，Gao 等人进行了从在线论坛提取问答知识对的工作，提出了一种新的抽取方法，使用一种基于分类方法的序列模板进行问题检测，然后使用一种图的后向传播方法进行对该问题的答案进行检测。其中还用到了一些余弦相似度方法、极大似然语言模型和 KL 散度语言模型等方法。实验表明，该种方法取得了不错的效果[15]。同年，Ding 等人从在线旅游论坛中提取大量语料，利用条件随机场(Conditional Random Fields, CRFs)进行问答对的提取工作，由此提出了一种问答对提取的通用框架。同时与 SVM、C4.5 进行了对比，并针对在线论坛语料的一些特点，进行了 chain CRFs 和 2D CRFs 的改进，最终 CRFs 图模型的效果最好[16]。2009 年，Cao 等人在 Shilin Ding 等人的研究基础上考虑了个性化因素，处理上下文关联和问题的答案信息的提取。提出了一种新的图表示方法，定制了结构化 SVM 方法，同时可针对不同的任务定制损失函数，效果不错的同时具有一定的灵活性[17]。同年，哈工大 Wang 等人首次进行了对在线论坛中文问答匹配对的提取工作。采用基于序列规则的方法发现问句，然后利用基于论坛结构的非文本特征提高答案检测的效果。文中使用到了规则、余弦相似度、基于 HowNet 的语义相似度、支持向量机等技术，同时指出了中文论坛中用户表达多用短句，与表达方式与表达习惯相关，中文语义更加多样复杂，对于语义理解存在一定的困难[18]。2010 年 Heilman 等人使用依赖性解析树中的最小编辑序列进行匹配[19]，2013 年 Severyn 等人利用在解析树上的判别树编辑的特征提取和工程进行匹配[20]， Yih 等人在 WordNet 上构建语义特征[21]。

可以发现，以往自动问答相关的工作主要是用规则策略、句法语法解析、语义词典、特征工程以及传统的浅层机器学习模型来完成的，但研究者们都是朝着人机自然交互以及返回唯一准确的答案的目标而努力的。

1.3.3 基于深度学习的自动问答技术研究现状

近几年来，随着深度模型的兴起和数据量的庞大，很多研究者会选择使用深度学习的模型去完成问句匹配、答案选择等自动问答相关任务。这些方法一般用以下的方式来解决问题。（1）在问题和答案上构建联合特征向量，然后将问题转化为一个分类或排序问题去解决[22][23]；（2）2015 年 Bahdanau 和 Vinyals 等人提出的文本生成的深度模型[24][25]；实质上也可以用到答案选择和生成上。（3）问题和答案的代表就可以通过某些相似度衡量去学习和匹配，使用类似思路，2015 年 Feng 等人用深度模型进行答案选择[26]，dosSantos 等人学习混合表示去检索语义上等价的问题[27]。2015 年，ACL15 上，李航老师团队在微博数据

上使用单层 LSTM 的 seq2seq 模型，并添加 attention 针对单轮对话做了生成，每句限长 140 个字。每条微博有平均 20 条回复[28]。与之不同的是，2015 年，周等人对多轮对话使用双层 LSTM 的 seq2seq 模型生成对话的表示，并用此做了对话的主题识别[29]。

此外，2016 年 Rocktäschel 等人还开发了一个在前提与假设之间的双向 attention 机制[30]。2016 年，Watson 核心技术团队 Ming Tan 等人使用了 CNN 和 RNN 相结合的深度模型处理答案选择问题，利用 CNN 主要侧重于 N-gram 局部相关信息和 RNN 主要侧重于长范围的信息而忽略局部细节信息的优点，在 TREC-2015 和 InsuranceQA 的公开集上进行了实验，同时也提出了一个为长答案序列建模的简单而高效的 attention 机制[31]。

目前，在问答领域比较流行的解决方案是 seq2seq 的模型，尤其是构建一个任务导向(task-oriented)的交互式问答机器人，比如订票，客服等等，是一件具有挑战性的事情。不少研究者对于此也做了大量的研究。2016 年，Li 等人采用两个机器人相互对话的方式进行，将 seq2seq 与 RL 整合在一起，可以使得回答生成时考虑得更加长远[32]。Tsung-Hsien Wen 等人将具体的业务信息和历史信息加到了模型中，结合 seq2seq 的句子建模，对话状态追踪，知识库，生成模型最终生成回答[33]。

1.4 本文主要研究内容

本文的题目为交互式问答中的语句关系识别方法，包括对交互式问答数据中两种重要关系匹配关系和补充关系的识别，主要研究内容包括以下几个方面：

第一，详细分析客服对话数据集中这种多轮对话的数据特点，包括对话中句子的关系、种类，对话的跨度等等，无论是在实验还是实际使用中，这些数据集的特点不仅对数据的构造有指导意义，同时为后期实验的设计与对比提供数据分析基础。

第二，研究基于卷积神经网络的和基于循环神经网络的两种不同类型的深度网络结构分别在这两种数据的问答匹配关系识别上的适用性，进行实验对比和性能分析。探索社区问答数据和客服对话数据两种数据的数据特点以及它们之间的差异。

第三，研究基于 RNN 的生成模型在对话数据上问答匹配的效果，对不同的数据组织方式，不同的 RNN 网络结构，包括添加机制，网络循环方向，问句的完整性、问句和答句的信息贡献度等进行实验对比分析，探索其性能表现的事实根据和解释，研究 RNN 相关网络结构对于对话中问题或者答案信息自动归纳与生成的有效性。

第四，针对电商客服对话数据，研究基于 RNN 的端到端的模型生成的信息对于答案或者问题的分类的影响，针对包括不同网络结构、不同的对话数据抽取方式、问句完整性在阈值的选择上进行分析。

第五，研究对话数据中，语句补充关系的识别，在进行答案自动归纳时，我们期望问句能够给出完整的信息，而人工对话数据集中，有一部分数据是存在补充关系的。本文试图使用支持向量机、卷积神经网络和串联循环神经网络对补充关系的识别进行了研究，分析不同模型、不同词向量在人工客服对话数据上，补充关系分类的效果，最终将补充关系识别与匹配关系识别相结合应用于问答匹配测试。

1.5 论文的组织与结构

本文主要探究交互式问答中问答匹配关系的识别方法和语句补充关系的识别方法，具体的章节组织结构如下：

第 1 章，是本文的绪论部分，将主要描述本课题的背景、研究目的和意义，以及与本课题相关的一些产品、评测和技术的研究现状。此外对本文涉及的主要研究内容也将进行详细的阐述。

第 2 章，将介绍本文中使用的相关技术，主要包括语义匹配的浅层机器学习方法、问答匹配的深度学习模型，包含有支持向量机，卷积神经网络，循环神经网络中的长短期记忆网络和门限循环单元，词向量技术、注意力机制，以及无监督文本相似度计算方法词移动距离。

第 3 章，将针对对话数据中问答匹配关系识别的两种数据抽取思路和问题进行阐述，针对人工客服对话数据进行详细的统计分析。同时，对社区问答数据和对话数据的预处理和抽取进行介绍，对词向量的训练和使用到的词向量进行介绍。

第 4 章，针对问答匹配任务，将介绍实验中对比方法的实现细节，包括基于 CNN 的方法和基于 RNN 的方法，同时还包括注意力机制以及不同的循环方向。进行不同类型数据，不同抽取方式的数据，不同方法的对比实验，介绍实验环境和实验评价指标，分别对排序结果和分类结果进行评价和分析。

第 5 章，针对补充关系识别，将介绍不同的实验方法，包括支持向量机，基于 CNN 的模型和基于 LSTM 的模型。并对实验数据的处理进行介绍。针对不同方式处理的数据、不同方法、不同词向量进行对比实验，并对结果进行分析。最后，阐述将补充关系识别方法和匹配关系识别方法相结合，进行问答匹配测试。

第 2 章 问答语句关系识别相关技术介绍

2.1 引言

本章主要介绍本文在问答语句关系识别中所使用到的一些浅层机器学习方法、深度学习模型以及与深度模型相关的机制等、文本相似度计算方法。包括支持向量机、词向量模型、卷积神经网络、循环神经网络中的长短期记忆网络、门限循环单元、词移动距离等。

2.2 浅层机器学习方法相关技术介绍

传统的语句关系识别技术主要包括基于规则的模型，基于句法语法解析树的和基于特征的浅层机器学习识别方法。这里主要介绍本文中使用到的一种基于特征的有监督浅层机器学习分类方法——支持向量机(Support Vector Machine, SVM)[34]，其广泛应用于模式识别、分类等任务。最先是由 Vapnik 等人对线性分类器提出了最佳设计准则，后来经过大量的研究，Boser 等人提出将内核技巧应用于最大余量超平面来创建非线性分类器，这就是今天支持向量机的标准方法。针对于语句关系识别而言，需要构建语句匹配对的特征空间，利用支持向量机进行特征空间的划分来对语句关系分类。

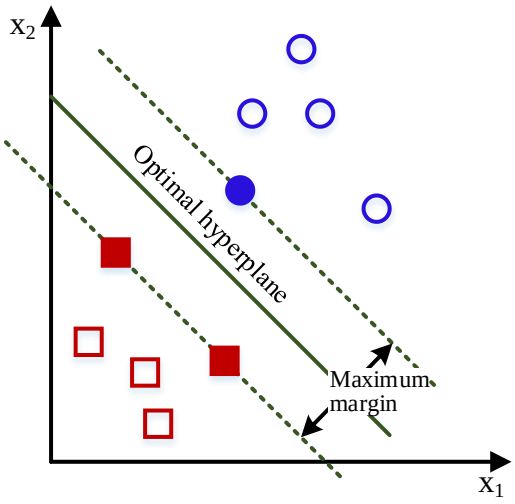


图 2-1 SVM 分类原理图[34]

支持向量机是定义在特征空间上的间隔最大的线性分类器，其分类的简单示意图如图 2-1 所示。SVM 的核心关键点主要在于升维，线性化和松弛变量。由于原本的 SVM 是线性分类器，针对在数据原本的特征空间中不可分的情况，

其主要手段是通过将低维特征空间映射到高维空间，然后找到一个最大间隔超平面对更高维的特征进行划分。这里使用到了核函数的技巧，将点积替换为非线性核函数，对低维特征进行升维。虽然在变换后的高维空间中是线性的，但该核函数的变化是非线性的，这就保证了其在原始空间中的分类时的非线性，同时也保证了在高维空间中转换为了线性，运算不像原来那么复杂。其中，部分核函数的计算方法有如下：

（1） 线性核函数(linear kernel)：

$$k(x, x_i) = (x \cdot x_i) \tag{2-1}$$

（2） 多项式核函数(polynomial kernel)：

$$k(x, x_i) = (s(x, x_i) + c)^d \tag{2-2}$$

（3） 径向基核函数(radical basis function)：

$$k(x, x_i) = exp(-\gamma \left| x - x_i \right|^2) \tag{2-3}$$

（4） Sigmoid 核函数(Sigmoid tanh)：

$$k(x, x_i) = \tanh(s(x, x_i) + c) \tag{2-4}$$

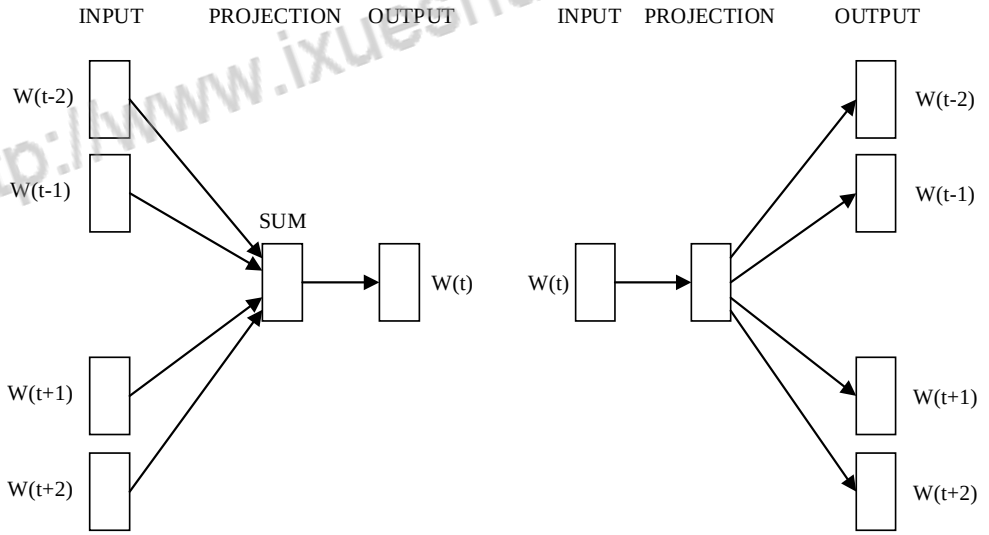
而对于松弛变量，主要是针对向高维空间中映射时线性不可分的以及样本中存在噪音的数据。简单的说，在某些特定情况下，我们无法严格的在样本数据点中找到一个分类的超平面，此时，允许模型具有一定的容错性，忽略一些离群点对模型精确度的影响，使部分样本点落在了超平面的间隔范围内。在实际的使用中，我们需要提取样本的多维特征，将特征矩阵和类标矩阵送入支持向量机进行模型训练以预测样本。

2.3 深度学习方法相关技术介绍

与浅层的机器学习方法相比，深度模型具有自动学习原始数据特征的能力，避免了大量的特征工程，同时也避免了需要大量专业知识相关的人力和物力，只要能够向量化表示数据，该种类模型就可以被快速的应用到其他方面，具有较高的移植性，无论是对于学术界的研究还是工业界的产出，这都是一个比较好的特性。同时，深度学习通过多层的非线性运算能够更加贴切的表征数据的特点，学习到数据的高层抽象特征与表示，在机器学习的很多任务上都取得了显著的效果。由于本文所处理的是自然语言相关的文本数据，它是一种人类大脑的高级活动，具有无法直观发现的抽象特征，而通过深度神经网络的学习可自动表征这种特征。基于以上原因，本文主要使用了深度学习相关的不同模型进行了对比，以下对深度模型相关技术进行了介绍，包括词向量模型、句子向量表示模型、卷积神经网络模型、循环神经网络模型和 attention 机制等。

2.3.1 词向量模型

词嵌入最早是由 Bengio 等人提出的一种具有语义特征信息的词的向量化表示，他们使用的是一种基于输入层、线性映射层、非线性隐藏层及 softmax 输出层的四层前馈神经网络[35]对词向量进行学习，Bengio 将训练出的词向量称为分布式表示(distributed representation)。而以往的 one-hot 词向量采用的是词袋(Bag of words, BOW)模型，其维度与词汇表的维度一致，在对应词的位置上分别用 0 和 1 表示该词是否出现。这种词向量具有维度高、稀疏的特点，可能导致“维数灾难”。另一方面，它并不能很好的表征词语中的语义关系，在进行语句相似度的计算时，容易带来错误。不同于词袋模型，Bengio 等人提出的词向量也叫作词嵌入，是一种低维实数向量表示，不仅克服了维数灾难，同时也能够学习到词与词之间的语义关系。后来，Mikolov 又在这方面做了一系列研究，于 2013 年提出了 CBOW(Continuous bag of words)和 Skip-Gram 两种语言模型[36]对词向量的训练进行建模，这是如今最常用的两种训练词向量的模型，两种模型无本质区别，只是 CBOW 是给定上下文预测中间的词，而 Skip-Gram 是给定一个词预测上下文。其基本结构如图 2-2 所示，W_*表示词，INPUT 表示输入层，PROJECTION 表示隐藏层，OUPUT 表示输出层。



(a) CBOW 模型[36]　　(b) Skip-Gram 模型[36]

图 2-2 词向量模型原理图

本文中使用的是 CBOW 模型，CBOW 是一个三层神经网络，包括输入层、线性隐含层、和输出层，输入已知上下文输出对下个单词的预测。首先，根据语料库建立词汇表，词汇表中所有单词拥有一个随机的词向量。然后，将单词 W 的上下文的词向量输入 CBOW，由隐含层累加求和，在输出层沿着哈夫曼树

的某个路径到达叶节点，预测单词 W，同时也确定了路径上所有分类器应作出的预测。采用梯度下降法调整输入的词向量，使实际路径向正确路径靠拢，训练结束即可从词汇表中得到每个单词对应的词向量。

2.3.2 CNN 模型介绍

卷积神经网络(Convolutional Neural Networks, CNN)是深度学习模型中一种常见的网络架构。Hubel 和 Wiesel 通过研究猫的视觉皮层细胞，提出了感受野的概念，受此认知机制的启发，1984 年 Kunihiko Fukushima 提出了 CNN 的前身——神经认知机(neocognitron)。20 世纪 90 年代，LeCun et al.等人发表论文，确立了 CNN 的现代结构，并对其进行了完善，被称为 LeNet5。直到 2012 年，Alex Krizhevsky 发表了 Alexnet[37]，扩展了模型的训练深度，在 GPU NVIDIA GTX 上完成了模型训练，为之后深度学习大体量、深层次的计算和训练提供了可能，将 LeNet 的思想扩展到了更大的能学习到远远更复杂的对象与对象层次的神经网络上。发展至今，卷积神经网络在视觉识别、语音识别和自然语言处理相关领表现优越。尤其是数据的涌现和 GPU 的发展，深度神经网络的研究成果越来越突出。

卷积神经网络的核心操作是卷积和池化，其他主要思想还包括局部连接和权值共享，实际实现中，我们可以进行层次的卷积和池化来达到提取特征的目的。卷积神经网络除了在图像任务上表现优越，在自然语言文本任务上也取得了不错的效果。如图 2-3 所示，是卷积神经网络对句子建模分类的示例[38]。

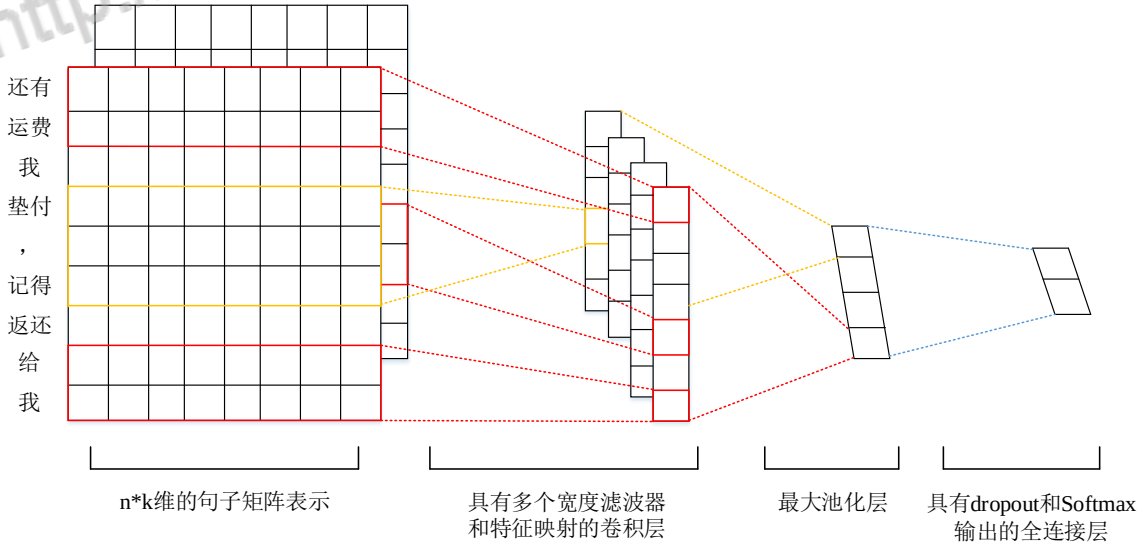


图 2-3 卷积神经网络句子建模分类示例[38]

以句子“还有|运费|我|垫付|，|记得|返还|给|我”为例，句子中的每个词用一个 k 维向量表示，句子有 n 个词，则该句子就表示为一个 n*k 维的矩阵，作为

卷积神经网络的输入。假设卷积窗口为 h，那么特征c_i则由公式(2-5)计算完成。

$$c_i = f(Wx_{i:i+h-1} + b) \tag{2-5}$$

其中，f是一个非线性激活单元，比如 tanh、ReLU 等。则该卷积所对应的特征映射为$\mathbf{c} = [c_1, c_2, \cdots, c_{n-h+1}]$。如果进行多个卷积核操作，我们将得到多个特征映射矩阵，在句子建模中，我们也可以选择做多种类型的 n-gram 卷积。在对特征矩阵执行卷积操作时，所有的滤波器组共享相同的权值。而这种卷积操作，表示了在该特征矩阵上，存在局部高度相关的。池化层一般有最大池化和平均池化，这里我们以最大池化为例，计算方式如公式(2-6)所示。其核心思想就在于抓住特征映射中最重要的部分，合并相似的特征信息，同时也承认了无论是在图像还是在文本信息上，均存在信息的不变性和可平移性。

$$\hat{c} = \max(\boldsymbol{c}) \tag{2-6}$$

经过拼接，得到一次卷积池化后的句子特征矩阵。当然，可以根据需要，对特征矩阵执行多次卷积池化操作，最终生成一个信息高度集中的维度较低的特征矩阵，既抽取了特征，又降低了维度。对于卷积神经网络的训练，依然采用反向传播的方式，使用随机梯度下降的方法。

2.3.3 RNN 模型介绍

鉴于全连接的 DNN 无法对时间序列上的状态变化进行建模，而在许多文本、图像、语音数据和任务上，比如机器翻译、对话处理上，均包含时间序列信息的特点。由此，循环神经网络(Recurrent neural networks, RNNs)应运而生。循环神经网络是一种能够通过节点循环抓取序列动态的连接网络。与标准的前馈神经网络不同，循环神经网络能够保持表征任意长上下文窗口中信息的状态特征[39]。正是由于循环神经网络循序其内部节点的循环，其能够选择对序列信息的持久化，基本结构如图 2-4 所示。

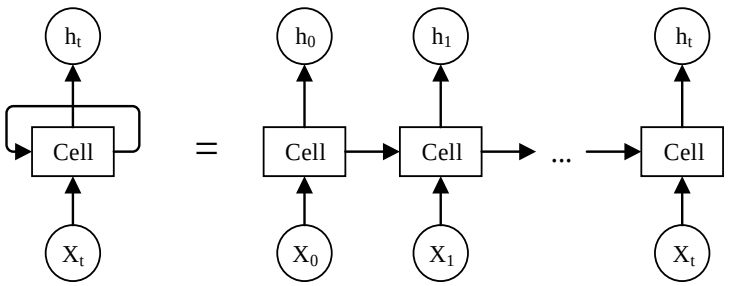


图 2-4 RNN 基本结构[1]

[1] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

从图中可以看出，神单元每次均接收上一状态的输入和当前输入，这就保证了时序信息能够从当前时刻传递到下一时刻，这就是它能够处理时序的关键所在。当然，Cell 只是一个总称，其内部计算方式，各有不同。由于基本的循环神经网络并不能很好的捕获到长程依赖的信息[40]，同时也可能会产生“梯度消失”问题，为了克服这些缺点，后来又出现了其他 Cell 的计算方式，其中比较有名的有长短期记忆网络(Long Short Term Memory networks, LSTM)[41]、门限循环单元(Gate Recurent Unit, GRU)[42][43]等。长短期记忆网络单元有三个主要的门，分别是遗忘门(forget gate)、输入门(input gate)、输出门(output gate)，通过不同门的计算，可以选择舍弃什么信息和添加什么新的信息。由 Cho.et.al 等人提出的门限循环单元是目前比较流行的 LSTM 的一种变体，在将重要特征保留的同时减少了参数运算，速度更快。其基本结构如图 2-5 所示。

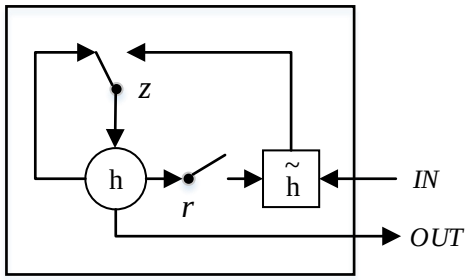


图 2-5 GRU 原理图[43]：r 为重置门，z 为更新门

GRU 将 LSTM 的单元状态和隐藏状态合并，将原来的三个门替换成了重置门和更新门。重置门用来决定新的隐藏状态$\tilde{h}_t$中上一隐藏状态$\tilde{h}_{t-1}$所占的比重，新的隐藏状态$\tilde{h}_t$由重置门过滤后的h_{t-1}和输入x_t经过非线性函数 tanh 激活得到。更新门用来决定当前隐藏状态h_t中新的隐藏状态$\tilde{h}_t$所占的比重，控制输出状态h_t中新的隐藏状态$\tilde{h}_t$和前一输出状态h_{t-1}所占的比重。

2.3.4 句子向量表示模型

句子向量表示是自然语言处理中的一个基础性工作，当我们需要做语句相关的任务时，首先需要向量化表示该语句，然后再进行其他模型的运算。以往，对于句子的向量表示主要使用的是词袋模型，这种模型存在很大的弊端，仅仅将文本看作是一个等概率的词集合，忽略了文本中的词序、语法和句法信息，导致语句相似度计算的不准确。后来，随着深度学习神经网络的兴起，文本的向量表示方法也丰富起来，包括卷积神经网络，递归神经网络，循环神经网络等。本文中主要使用了卷积神经网络和循环神经网络对句子进行建模，这里仅对这两种方法作简要介绍，具体会在后面实验方法章节详细介绍。

方法一 循环神经网络句子建模

循环神经网络主要是通过循环输入当前词向量和上一个神经元的状态得到当前神经元的输出。以最初始的循环神经网络单元为例，假设给定一个句子，其词向量序列为$(x_0, x_1, \cdots x_n)$。那么，在 t 时刻神经元的输出状态s_t的计算公式如(2-7)所示。

$$s_t = f(Ux_t + Ws_{t-1} + b) \tag{2-7}$$

式中 U，W，b——表示参数；

x_t——表示当前词向量；

s_{t-1}——表示前一时刻的状态输出；

f——表示激活函数。

这里的 f 一般是非线性的激活函数，如 sigmoid、tanh 等，这里的参数 U，W，b 在每一层是共享的，最终时刻 n 的输出状态s_n即为句子的向量表示。

方法二 卷积神经网络句子建模

卷积神经网络的核心操作是卷积和池化，首先，通过词向量的表示方式将文本表示为一个矩阵，然后送入卷积神经网络进行卷积和池化操作，根据实验的效果可以选择卷积和池化的次数，通过这种方式能够自动从句子中提取特征，获取包含语义特征的句子向量表示。

2.3.5 Attention 机制

注意力机制是受人类的视觉注意机制启发而提出来的，通俗的讲，当人类去观察一幅图片时，会不自觉的聚焦到图片的重点信息或者自己需要挖掘的信息上而忽略其周边的信息。事实上，就是按照“高分辨率”聚焦在图片的某个特定区域并以“低分辨率”感知图像周边区域的模式，然后不断地调整焦点。最早，这种机制应用于视觉图像领域，其中真正流行起来的是 2014 年，Google mind 团队在图像分类任务上使用了 attention 机制[44]。同年，Bahdanau 等人使用类似 attention 的机制在机器翻译任务上将翻译和对齐同时进行[24]，首次将注意力机制应用到了自然语言处理任务中，这里应用于 encoder-decoder 模型。随后，注意力机制又被扩展应用于卷积神经网络模型。编码解码模型，是建立在输出矩阵包含句子中全部的特征信息的基础上进行的，事实上实验表明，RNN 编码出的特征矩阵并不能包含全部的信息，尤其是在句子较长的时候，越靠前的信息越容易丢失，这一点在添加倒序的 RNN 编码信息有助效果提升上就有所体现。此时，加入注意力机制，就能自动更新权重，动态的选择重要的特征信息添加到编码的特征矩阵中。以双向递归网络的 attention 为例，如图 2-6 所示。

y 是解码器生成的译文词语，x 是原文的词语。从中我们可以看到，每个解码器输出的词语y_t由所有输入状态的权重组合和最后的输入状态计算而得，并不仅仅由当前最后输出状态决定，其中，a 的权重值决定每个输入状态对输出状态的贡献度。通常，a 的求和结构是归一化的，表示不同输入状态的概率分布。比如，在生成词语y_t时，$a_{t,3}$的权重最大，这说明在生成该译文词语时，最需要关注的是x_3这个词语。同样的，在本文的课题中，句子上下文之间也存在重点信息，和需要减弱关注的信息，可以将类似的思想运用到句子的上下文之间。

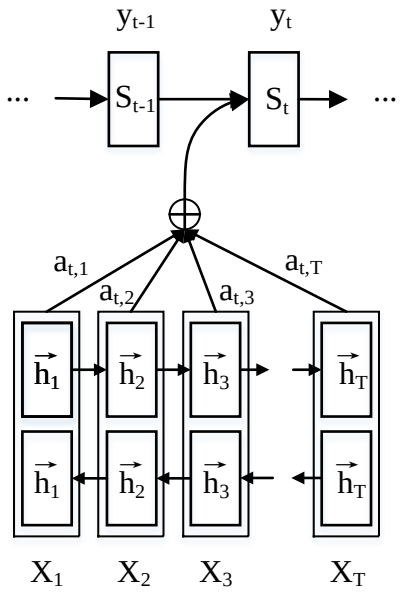


图 2-6 递归神经网络 attention 模型示例[24]

2.4 文本相似度计算方法

文本相似度计算方法有很多，这里主要介绍本文中使用到的词移动距离(Word Mover's Distance，WMD)，它是一种无监督的文本相似度计算方法，需要在词向量的基础上进行文本相似度计算。首次被提出，是在 Matt J 等人发表的论文中[45]。该方法在无监督文本相似度计算方法中，是一种计算效果比较好的方法。该方法借鉴了搬运的思想，巧妙地将文本相似度的计算问题转化为线性规划中运输问题的最优解问题。不难看出，词移动距离衡量的是句子 A 和句子 B 之间的不相似性，距离值越小表示两文本相似度越大。

以句子 A：“Obama speaks to the media in Illinois”和句子 B：“The president greets the press in Chicago”为例，其分布如图 2-7 所示。

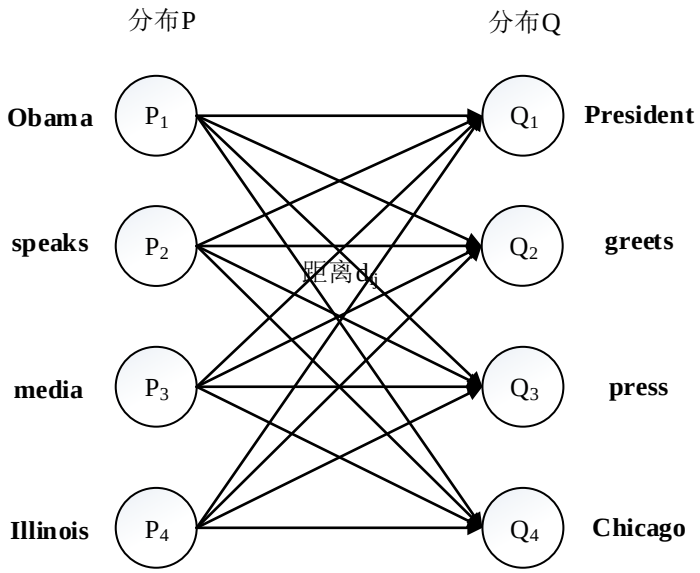


图 2-7 WMD 句子分布示例图

每个分布中，每个词以词向量表示，以词频为权重，分布之间每两个词之间计算欧式距离构成转移矩阵，两句子之间的距离就是在此约束条件下求解分布 P 准换到分布 Q 的线性规划最优解。具体计算方法如公式（2-8）所示。

$$\begin{aligned} &\min_{T\geq 0}\sum_{i,j=1}^{n}T_{ij}(i,j) \\ &subject\ to:\sum_{j=1}^{n}T_{ij}=d_i \quad \forall i\in\{1,\cdots,n\} \\ &\sum_{i=1}^{n}T_{ij}=d_j^{'} \quad \forall j\in\{1,\cdots,n\} \\ &d_i=\frac{c_i}{\sum_{j=1}^{n}c_j} \\ &d_j^{'}=\frac{c_j}{\sum_{i=1}^{n}c_i} \end{aligned} \tag{2-8}$$

其中 $c(i,j)=\left\|x_i-x_j\right\|_2$ ——表示文档 A 中的词 i 到文档 B 中的词 j 的距离；

T_{ij} ——表示单词 i 移动到单词 j 的权重；

d_i ——表示单词 i 在文档 A 中的词频；

$d_j^{'}$ ——表示单词 j 在文档 B 中的词频。

2.5 本章小结

本章重点介绍了本课题中使用到的主要方法，具体有浅层的机器学习方法支持向量机，深度学习模型相关的技术包括词向量方法、卷积神经网络模型、循环神经网络中的长短记忆网络、门限循环单元，文本相似度的计算方法词移动距离，便于后期章节方法的阐述。

第 3 章 对话实验数据的分析与构建

3.1 引言

本章主要介绍针对对话数据问答对提取的整体思路和结构，以及如何解决从整个对话问答匹配中抽象出的匹配关系识别问题。同时，对数据进行了统计剖析，给出解决方案和数据构建的理论依据所在，并对数据构建方法以及模型训练前期的预处理工作和词向量训练工作做了介绍。同时，也对实验中使用的 cQA 数据的处理做了简单介绍。

3.2 问题描述与整体结构

本课题主要处理的是某电商客服的对话数据，这种数据最大的特点是每对客户和客服人员之间往往存在多轮对话的交互，这就导致了该种数据中存在两种重要的语句关系——匹配关系和补充关系。语句匹配关系指对话问答数据中客户问句与客服答句之间的匹配关系，语句补充关系是问句之间或者答句之间的补充关系，即后述语句对前述语句的补充说明。自然地，本文的研究就聚焦到这两种关系的识别。

与社区问答不同的是，我们无法明确界定对话的边界，同时除了一对一的匹配之外，对话数据中存在多对多的复杂的匹配形式，要想直接通过某个模型实现比较困难，我们采用分而治之的思想，将多对多匹配拆分为一对多的匹配形式。针对此种数据，我们提出了两种实验性的构想进行问答匹配相关的实验。如图 3-1 和图 3-2 所示，分别是基于每轮对话的问答匹配流程图和基于对话窗口的问答匹配流程图。具体的数据处理将在之后的章节详细阐述，本节主要阐述问题的解决思路并抽象出问题的本质。

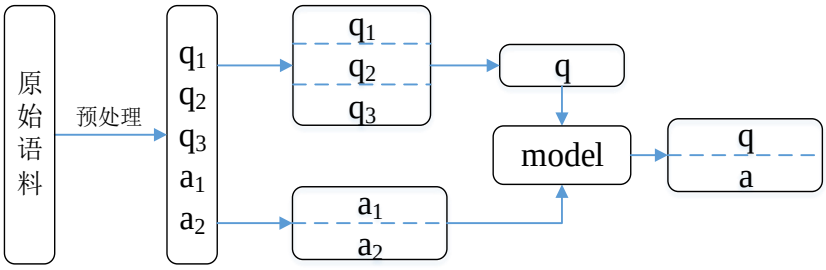


图 3-1 基于每轮对话的问答匹配流程

图 3-1 中，q_*表示问句，a_*表示答句，$(q_1, q_2, \cdots, a_2)$表示原始语料中抽取出的一轮对话，(q_1, q_2, q_3)中可能存在正确的问题，也可能存在不处于当前轮对话

的问题，q 表示从(q_1, q_2, q_3)中抽取或者学习到的正确问句。基于每轮对话的匹配，是在每轮对话中，假设问句给定的情况下，从下文中抽取出与该问句语义匹配的答案，也就是图 3-1 的 model，这里的问句是图 3-1 中的 q，这里的下文是指包括(a_1, a_2)在内的从q_3开始的 n 条下文，这里的根据数据统计与分析结果而定，会在 3.4.2 节进行详细说明。

由此，实验问题便抽象为：给定一个客户问句和多条下文$(q, c_1, c_2, \cdots, c_n)$，对下文做答句排序和选择的任务，即问句与多条下文之间的语句匹配关系识别。

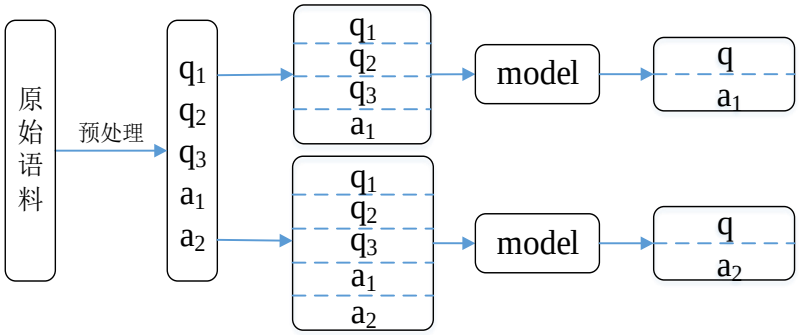


图 3-2 基于固定对话窗口的问答匹配流程

图 3-2 中，q_*表示问句，a_*表示答句，$(q_1, q_2, \cdots, a_2)$表示原始语料中的一个对话片段，在该对话片段中，以答句为基准进行数据组抽取，$(q_1, q_2, \cdots, a_1)$表示以a_1为基准向上抽取其 n 条上文，$(q_1, q_2, \cdots, a_2)$表示以a_2为基准向上抽取其 n 条上文，这里的 n 经过实验数据统计分析确定，上文中可能包括匹配的问句、不匹配的问句、答句，q 是通过模型 model 从上文中学习到的与基准答句a_*匹配的问句。基于对话窗口的问答匹配，不区分对话轮，以客服的答句和固定的窗口大小为基准，切分对话片段，做问答匹配，也就是图 3-2 中的 model。

由此，实验问题便抽象为：给定一个客服答句和多条上文$(c_1, c_2, \cdots, c_n, a)$，对上文做问句排序和选择的任务，即答句与多条上文之间的语句匹配关系识别。

区别于常规的检索和单对单的匹配，交互式问答数据中存在丰富的场景信息，具有上下文相关的特征，简单的通过规则或者浅层机器学习的方法，生成语义上匹配的问答对，难以实现对此种类型数据的处理。通过深度模型的训练，自动学习上下文语义关联的特征，能在一定程度上解决此类问答匹配。

3.3 数据评估与分析

这里主要对人工客服对话数据进行详细的统计和分析，这代表了部分对话数据的特点。这是我们子问题的提出、实验数据的构造和实验解决方案的设计的重要依据。

我们统计的指标包括人工客服数据中，语句间匹配关系和补充关系的占比，

每轮对话的匹配类型的占比，不同关系个数的对话的占比，每轮对话跨度的占比。每轮对话的匹配类型的占比，分为两大类，一对一匹配和包含补充关系的多对多匹配。匹配类型的占比，表示我们处理的问答匹配任务的类型。这些结果跟我们处理匹配关系识别和补充关系识别紧密相关。

表 3-1 匹配关系和补充关系的占比

语句关系	匹配	补充
占比	61.26%	38.74%

从表 3-1 中，可以看出，整个对话数据中语句之间的关系主要包括匹配和补充两种关系。匹配指的是客户问句和客服答句之间的匹配关系，补充指的是在同一轮会话中客户问句之间的补充关系和客服答句之间的补充关系，以问句为例，问句补充关系指几个问句表达相同的意图，组合在一起完整的表达客户的问句意图。从表中可以看出，匹配关系占到了 61.26%，补充关系占到了 38.74%，体现了对话数据中需要解决的两大子问题，本文将会在后面的章节针对这两个问题作详细阐述和建模。

表 3-2 对话不同匹配类型占比

匹配类型	一对一	多对多
占比	59.40%	40.60%

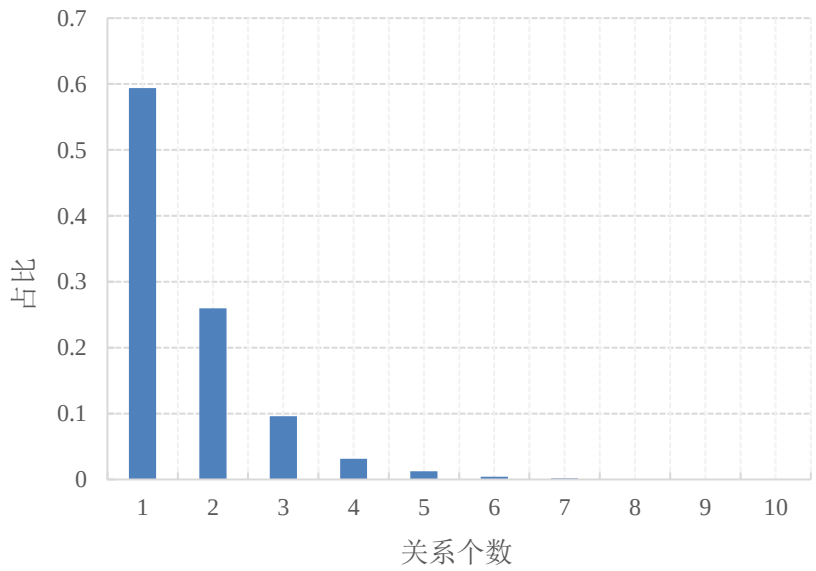


图 3-3 不同关系个数的对话占比

从表 3-2 中，我们不难发现，匹配关系中一对一匹配的占比达到了约 60%，这是其最基本的关系，也是有重要影响的一个匹配类型。再则，从图 3-3 中包含 2 种关系的对话的占比发现，一对多的关系占到了 25.97%，而所有的多对多

的关系总计才占 40.60%，也就是说多对多的关系中 63.97%的一对多的关系，因此，除一对一的匹配关系外，一对多的匹配关系是第二重要的问答关系类型。不难看出，这两种匹配关系的占比达到了约 85%。源于此统计依据，我们的模型重点解决的就是这两种类型的问答匹配。

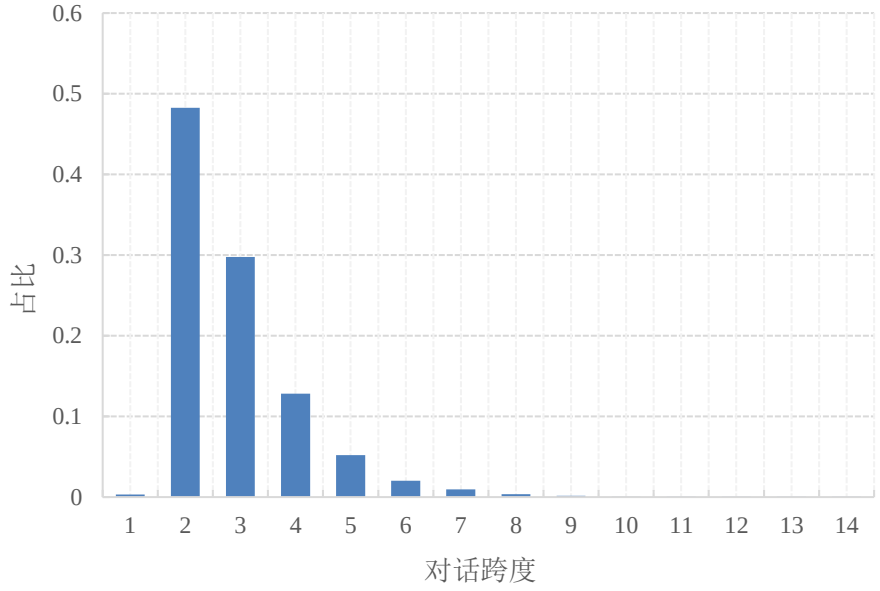


图 3-4 不同跨度的对话占比

图 3-4 中，统计了不同跨度对话的占比，其中跨度小于等于 6 句的对话占比大约为 98%。这表现了我们对话数据中一个重要特点，从客户给出问题到客服人员回答完毕，其对话语句数基本不超过 6。这也就决定了我们在进行匹配关系识别实验抽取每轮对话时，将会选择跨度区间在[2,6]之间的数据，同时后期生成模型组别大小的设定也会参考此跨度范围。

3.4 数据预处理与构建

本节主要对数据的处理以及实际输入到模型中的实验数据的构造作详细阐述。为了同时检验我们的模型在社区问答数据和对话数据上的效果，以及不同模型在不同类型数据集上的表现，我们分别选取了一份社区问答数据和一份某电商人工客服对话数据作为实验数据。

3.4.1 cQA 数据处理与提取

在对比多个数据集的情况下，选择了 Semeval2016 的 cQA 数据集。该数据集来源于卡塔尔生活论坛上的真实数据，其原始数据格式如下图 3-5 所示。<RelQBody>表示完整的问题，<RelCText>表示该问题对应的评论，

<RELC_RELEVANCE2RELQ>表示评论与问题是否相关，每个问题下对应的有10条与之相关或者不相关的评论。在官方所给的标注数据中，相关度分为3类：“Good”，“Potentially Useful”和“Bad”。实际处理时，我们会将标注为“Potentially Useful”和“Bad”的都视为不相关。

```
<Thread THREAD_SEQUENCE="Q268_R16">
      <RelQuestion RELQ_ID="Q268_R16" RELQ_CATEGORY="Moving to Qatar"
      RELQ_DATE="2013-07-31 02:27:08" RELQ_USERID="U5151"
      RELQ_USERNAME="shehabi">
          <RelQSubject>Best Bank.</RelQSubject>
         <RelQBody>Hi ti all QL's; What bank you are using? and why? Are you using this
            bank just because it has an affiliate at home? Regards;</RelQBody>
     </RelQuestion>
      <RelComment RELC_ID="Q268_R16_C1" RELC_DATE="2013-07-31 06:46:39"
        RELC_USERID="U65" RELC_USERNAME="Molten Metal"
        RELC_RELEVANCE2RELQ="Bad">
            <RelCText>banks are using us ... Talk to those who had taken a credit card or
              loan to know more ...</RelCText>
     </RelComment>
…
```

图 3-5 Semeval 原始数据格式示例

（1）数据抽取

由于我们使用了两种不同的深度学习模型，因此我们根据模型抽取了两种形式的数据。本节中以 q 表示问句，以 c 表示该主题问句下对应的评论。对于基于卷积神经网络的匹配模型，直接从原始数据中抽取<Q, C>形式的问答对即可；对于基于循环神经网络的生成模型，我们以组别大小为 6 来抽取数据，形式如<Q, C_1, C_2, …, C_n>。由于原始数据中是以一个问题 10 条评论的形式提供的，我们需要对数据进行重组。重组的规则如下，针对 10 条评论中的每一条正确评论，随机挑选其余 4 条与问题不相关的评论，将问题和这 5 条评论组合，构成一组基于循环神经网络模型的实验数据。选择组别大小为 6 主要与数据统计分析和模型有关。经过对该社区问答数据的简单统计发现，每个问题下对应的相关评论和不相关评论的比例不一，部分问题下存在较多的评论，部分问题下评论较少，同时评论语句与非完整答句不一样，不便于直接合并为一句，直接合并也可能会导致句子过长，而在社区问答实验中我们构建的循环神经网络的每组数据的标签是一个与该组数据中问题匹配的句子。因此，我们折中统一了组别大小为 6 抽取实验数据，既保证了实验数据中存在一条相关评论，同时也不会因为组别太大没有满足数量的评论，导致模型在后期补零的句子矩阵中多次进行无意义的计算。

（2）数据预处理

表 3-3 对预处理后的 Semeval 2016 数据的特点进行了说明，这里我们使用的是验证集和测试集，章节 4.6.1 对此做了详细说明。从表中可以看到，问题的平均长度大约在 40 至 45 之间，答案的平均长度大约在 30 至 35 之间，同时，我们对句子长度的不同占比也做了统计，大部分的句子长度不超过 200，最终确定了模型中句子的最大长度在 180。

表 3-3 SemEval 2016 处理后的数据特点

	Dev	Test
# of related ques.	244	327
# of cmt.	2440	3270
# of good ans.	818	1329
Avg. len. of ques.	41.94	43.40
Avg. len. of ans.	31.77	32.97
# of ques. w/o ans.	33	12

该社区问答数据中，存在一些对模型训练无意义的词，其中我们采用规则的方法对邮箱、网页标签、部分标点符号和特殊表情符号进行了过滤。使用 scikit-learn 中的相关工具对英文文本进行了 tokenize。处理后的样例，如表 3-4 所示，表中 Q 表示问题，C_*表示评论。

表 3-4 Semeval 2016 数据处理后的样例

类型	内容
Q	is there any place i can find scented massage oils in qatar
C_1	Yes It is right behind Kahrama in the National area
C_2	you mean oil and filter both
C_3	What they offer
C_4	What ’s the name of the shop
C_5	Swine No I do n't try with salesgirls My taste is classy

3.4.2 对话数据处理与提取

基于以上的数据统计分析，根据该电商客服对话数据的特点，对应于 3.2 节两种解决思路，本课题主要构造了两种对话实验数据，一种是基于每轮对话的，另一种不区分对话轮数，滑动固定窗口大小抽取数据。

这里标注的原始对话数据规模为 1975 组，过滤后为 1881 组，以比例 9:1 进行训练集和测试集的划分。主要过滤主题为“礼貌问候、闲谈（噪声）”的对话组，最终剩余 1881 组，选择跨度在 2-6 之间的每轮对话。对话标注数据样例如图 3-6 所示，每行数据主要有 6 列，分别表示对话 ID，句子 ID，说话角色，语句内容，句子补充关系，句子匹配关系。根据两种策略从中构建实验数据。

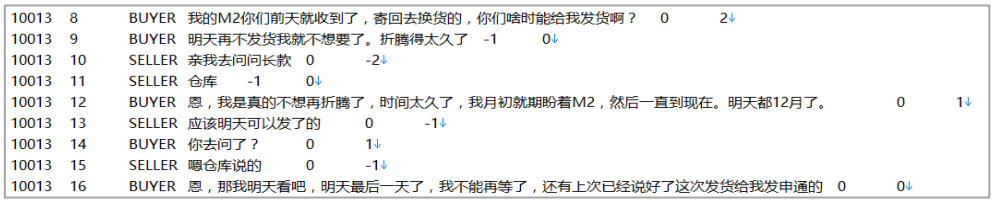


图 3-6 对话原始标注数据示例

策略一 从每组对话中抽取出每轮对话，具体样例如图 3-7 所示，每个红色方框内表示一轮对话。拼接具有补充关系的问题，组合后若组别不满 6 条语句，扩大窗口向下添加下文，直至满足组别长度为 6，无下文则保持当前组别数据长度。对于针对不完整问题的实验，将具有补充关系的每个问题分别与其所在对话轮中的下文进行组合，依然扩充对话轮至跨度为 6，与完整问题的对话轮扩充方式相同。同时，在符合条件的每轮对话中抽取由具有补充关系的答句拼接而成的完整答句作为该轮对话的类标。这里每轮对话是人工标注的结果，基于每轮对话的抽取保证了数据组中匹配问答对的主题一致性。最终抽取形式为<Q, C_1, C_2, …, C_n>，其中 n<=5，Q 表示问句，C_* 表示下文。具体算法描述如算法 3-1 所示。

算法 3-1 策略一数据抽取算法描述

输入： 所有 session 的对话
输出： 形式为<Q, C_1, C_2, …, C_n>的对话轮的集合

```
Begin
    遍历所有 Session：
    For 句子 Si in Session：
        If 句子 Si 是某轮对话中语义角色为 BUYER 的起始匹配句：
            抽取句子 Si 对应的对话轮和完整答案
            If 该对话为“闲聊”or“礼貌问候”or 对话跨度不在[2,6]之内：
                Continue
            抽取该轮对话中与具有补充关系的答句并拼接作为类标
            拼接具有补充关系的问句
                或将子问句分别和该对话轮中与完整问题对应的下文进行组合
                If 新的对话轮跨度小于 6：
                    若存在下文，添加下文扩充该轮对话跨度至 6
                保存每轮对话和该轮对话对应的答句类标
    End
```

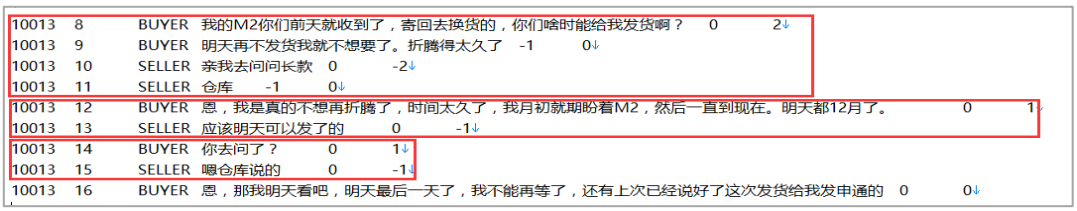


图 3-7 对话抽取策略一示例

最终抽取出的结果样例如表 3-5 所示：

表 3-5 策略一抽取数据示例

ID	内容
8、9	Q：我的 M2 你们前天就收到了，寄回去换货的，你们啥时能给我发货啊？明天再不发货我就不想要了。折腾得太久了
10	C_1：亲我去问问长款
11	C_2：仓库
12	C_3：恩，我是真的不想再折腾了，时间太久了，我月初就期盼着 M2，然后一直到现在。明天都 12 月了。
13	C_4：应该明天可以发了的
14	C_5：你去问了？

策略二 以窗口大小为 6，抽取角色为“SELLER”的 5 条上文，每 6 条数据为一组，若不满足 6 条，则维持当前组别长度，模型支持变长。同时，抽取出当前组别中，与“SELLER”句匹配的问题组合句作为类标，这里的匹配关系是经过人工标注所得，抽取出的数据组中匹配句对具有主题一致性。最终抽取形式为<$C_1, C_2, \ldots, C_n, A$>，其中 n<=5，C_* 表示上文，A 表示答案。具体标注数据样例如图 3-8 所示，具体算法描述如算法 3-2 所示。

算法 3-2 策略二数据抽取算法描述

输入： 所有 session 的对话
输出： 形式为<$C_1, C_2, \ldots, C_n, A$>的对话组的集合
1 **Begin**
2 遍历所有 Session：
3 遍历 Session 中跨度在[2,6]之间的对话轮：
4 遍历对话轮中语义角色为 SELLER 且与 BUYER 句匹配的语句 Si：
5 **If** 句子 Si 为非“闲聊”和“礼貌问候用语”：
6 抽取句子 Si 及其 5 条上文组合为一组
7 抽取该组别对话中与答句 Si 匹配的问句并拼接作为类标
8 保存每组别对话和该组别对话对应的问句类标
9 **End**

以下每个有色方框内为一组数据：

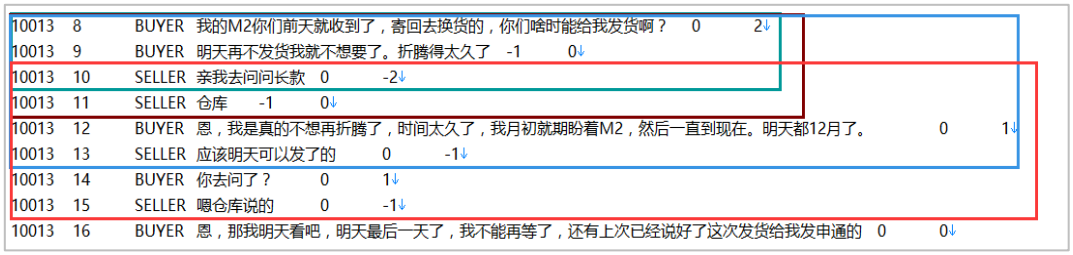


图 3-8 对话抽取策略二示例

抽取结果样例如表 3-6 所示：

表 3-6 策略二抽取数据示例

ID	内容
8	C_1：我的 M2 你们前天就收到了，寄回去换货的，你们啥时能给我发货啊？
9	C_2：明天再不发货我就不想要了。折腾得太久了
10	C_3：亲我去问问长款
11	A：仓库

3.5 词向量的构建

在文本上使用深度学习模型的前提是，必须将文本以矩阵的形式表示出来，词向量的训练就很好的达到了这样一种基础作用。由于词袋模型不能正确表达句子中词的词义、词序和语法信息，同时通过此种方式产生的词向量具有高维稀疏的特点，给模型的运算带来了巨大的压力。我们选择了由 Mikolov 等人提出的词嵌入的方式，将词表征为低维稠密的实数值向量。针对词向量的训练，这里我们选择了连续词袋模型(CBOW)作为训练词向量的模型，其结构已在 3.2 节介绍，不再赘述。

词向量的训练不需要人工标注的有监督数据，因此我们可以尽量扩充与实验数据相关的词向量训练语料，获取语义信息表达更丰富的词向量。这里我们主要选择了 wiki 英文语料，电商客服对话数据语料和百度知道问答通用语料。Wiki 英文语料和百度知道语料的规模比较大，达到了十多 G。电商客服的数据相对比较少。

处理之前以每行一个句子或者一篇文档的方式组织数据，进行词向量的训练。训练之前，需要先对数据进行分词，这里可以使用 scikit-learn，结巴分词等工具，然后再进行字向量或者词向量的训练。词向量维度选择上需要适中，当词向量的维度比较大时，能够更准却的表达词的语义信息，若词向量维度太大也会造成模型训练速度缓慢，引入了太过复杂的参数运算。这里我们选择 CBOW 模型进行词向量的训练，将滑动窗口大小设置为 5，词频小于 5 的词过滤，词向量维度设置为 200 维。最终训练得到了如下三类词向量，如下图所示，每行第一列表示词，剩余的列表示该词对应的词向量。

Wiki 英文语料词向量：包含 13967 词

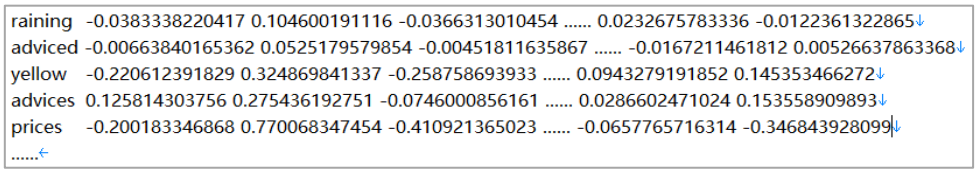


图 3-9 Wiki 英文语料词向量示例

客服对话数据词向量：包含 2748 词

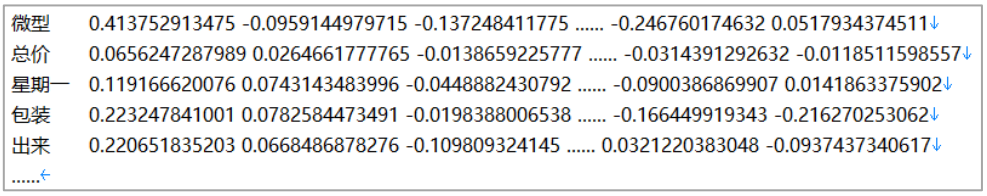


图 3-10 客服对话数据词向量示例

通用语料字向量：包含 8760 字

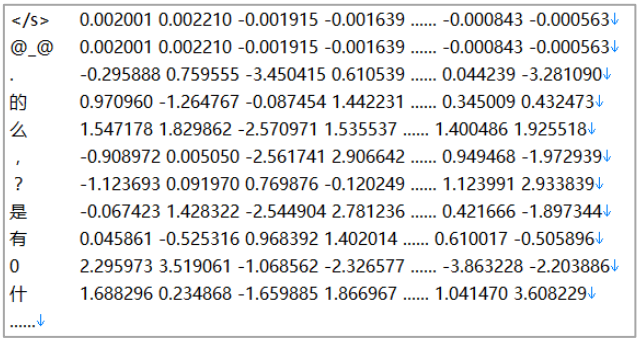


图 3-11 通用语料字向量示例

3.6 本章小结

本章主要对课题实验方案的整体结构以及课题所解决的问题进行了介绍，在数据统计和分析的基础上，阐述了解决方案制定和实验数据抽取方式的依据。同时，分别针对社区问答数据和人工客服对话数据，以及两种不同的的模型上，数据的预处理和抽取方式做了阐述，最后对输入模型训练的前提条件词向量的生成做了讲述。

第 4 章 问答匹配关系的识别

4.1 引言

本章主要介绍问答匹配模型主要使用的方法，由于深度模型具有自动学习特征的优势，在本章节的问题中，主要的解决方法使用的是深度学习相关的模型，包括基于卷积神经网络的问答匹配模型和基于循环神经网络的问题、答案信息自动归纳模型。之后，分别介绍了实验环境和评价指标，并对实验结果进行了详细的对比分析。

4.2 基于 CNN 的问答匹配关系识别模型

目前，在深度学习领域，卷积神经网络是应用最普遍的基本网络结构之一。其在多个领域多个任务上都取得了不错的实验效果，包括图像识别、语义匹配、情感分析等等，同时也有一部分运用到了实际应用中，比如很有名的 AlphaGo 计算机围棋。本文在使用基于卷积神经网络的匹配模型时，将问题或者答案的选择任务当作句子对间的语义匹配关系分类任务进行模型的构建，模型结构图如图 4-1 所示。图中，q 表示问句，a 表示答句。

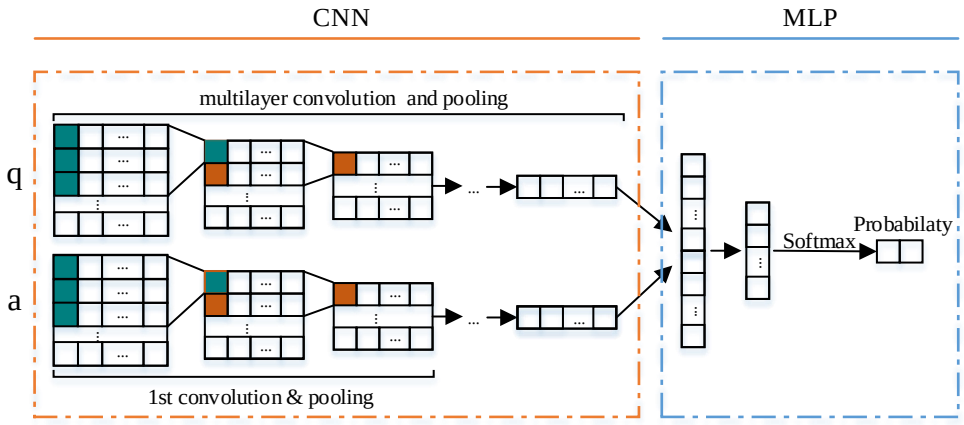


图 4-1 基于 CNN 的问答匹配关系识别模型结构图

通常，在一般的分类任务上，神经网络的顶层输入是一个数据矩阵，然后经过特征的学习后，经过多层感知器送入 softmax 进行分类。但是，在我们的数据中，要求的是针对一个问答句对进行打分，也就是两个句子给一个输出。因此，我们采用并行的卷积神经网络进行问答对的匹配。首先，通过词向量分别将两个句子 q 和 a 表示为两个大小为（句长*词向量维度）的二维矩阵。然后，将这两个句子分别作为卷积神经网络的输入，通过多次卷积和池化，对句

子进行建模，融合低级特征，抽取出蕴含丰富特征信息的高级特征矩阵。其实质，就是通过一系列的非线性运算，将一个高维特征矩阵，转换为一个包含特征信息的低维矩阵。经过多层次的卷积池化之后，将最后得出的特征矩阵拼接送入一个全连接的多层感知器，经过非线性映射后，使用逻辑回归分类器分类，计算极大似然估计的损失。

具体细节描述如下：

（1）输入处理

在我们实际处理的数据中，句子的长度是不一致的，而我们的模型要求输入的矩阵大小相同，这是使用卷积神经网络中用矩阵表示句子的一个普遍的问题，我们会强制设定一个句长，假设为 n。当句子的真实长度小于该设定值时，使用最多的一种做法叫作 Zero Padding，也就是对不足的做补 0 的操作。当句子实际长度大于设定句长时，将超出的部分截断。针对截断可能会有信息损失的情况，选择一个合适的句长值即可。针对补 0 的操作，我们比较担心会引入噪声。实际上，卷积是一个加权求和的过程，最大池化是一个选择最大信息的过程，经过这两步操作，会逐渐的将 0 过滤掉，保持原来最显著的特征信息。这种操作，既保证了输入的一致性，又不会遗失掉原有的特征。

（2）卷积层和池化层参数选择

卷积和池化是卷积神经网络模型的核心操作，其卷积池化的层数和卷积核参数的设定是比较重要的部分之一。理论上说，越深的网络训练出的模型，在测试集上的表现越好，不过往往也会出现过拟合的情况，同时训练也需要更大的时间成本。如何在网络深度、训练时间和训练效果之间找到平衡，是关键所在。由于分别在社区问答数据和客服对话数据上做了不同的实验，这里就这两种数据分别作卷积核和池化层的参数说明。与图像不同，我们更加注重的是句子字与字或者词与词之间的 n-gram 关系，因此这里我们采用了宽卷积的方式。

（a）**社区问答数据网络参数**　　句子矩阵输入是两个 180*200 的矩阵，训练过程中共进行了 3 层卷积和池化。第一层卷积和池化窗口大小为：filter_shape0=(3,1)，pool_size0=(2,1)；第二层卷积和池化窗口大小为：filter_shape1=(3,1)，pool_size1=(3,1)；第三层卷积和池化窗口大小为：filter_shape2=(3,1)，pool_size3=(27,1)。并行对句子对进行三层卷积池化后，最终得到一个 1*400 的特征向量，作为多层感知器的输入，经非线性变换后，送入 softmax 分类。

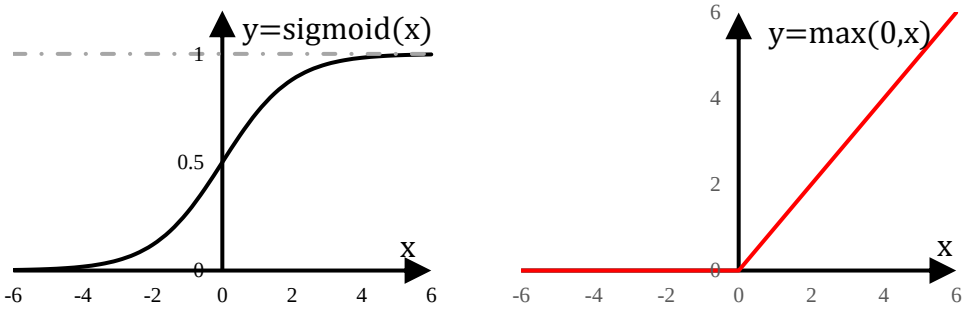
（b）**客服对话数据网络参数**　　句子矩阵输入是两个 30*200 的矩阵，训练过程中共进行了 3 层卷积和池化。第一层卷积和池化窗口大小为：filter_shape0=(3,1)，pool_size0=(2,1)；第二层卷积和池化窗口大小为：

filter_shape1=(3,1)，pool_size1=(2,1)；第三层卷积和池化窗口大小为：filter_shape2=(3,1)，pool_size3=(4,1)。并行对句对进行三层卷积池化后，最终得到一个 1*400 的特征向量。之后的计算方式与(a)相同。

大规模的神经网络在训练时往往比较费时同时还容易过拟合。为了防止过拟合，我们在卷积池化后，做了 dropout 的处理，随机的使一部分神经元失效。Dropout 实际就是每次训练时，随机让网络某些隐含层节点的权重不工作，相当于每个 batch 训练了不同的网络，通过不同模型防止了过拟合[47]。

（3）激活函数的选择

以往的卷积神经网络的激活函数多会选择 sigmoid，但它却存在着一定的弊端。如图 4-2 所示，随着 x 的增大，sigmoid 的函数曲线会逐渐平缓，导致导数趋近于 0，容易出现梯度消失。随着训练深入，模型的参数无法更新。另一种激活函数 tanh 与 sigmoid 相似。这里我们选择修正线性单元函数 ReLU 作为我们的激活函数，计算公式为 y=max(0, x)，函数图像如图 4-2 所示。从图中，我们发现，当 x 大于 0 时，随着 x 的增大，y 等比增大，斜率始终为 1，阻止了梯度消失的问题。



(a) Sigmoid　　(b) ReLU

图 4-2 Sigmiod 和 ReLU 函数图像

4.3 基于 RNN 的问答匹配关系识别模型

与常规的检索和语义匹配不同，对话数据中的问答匹配存在着丰富的场景信息，仅通过简单句对匹配，未必能够捕捉到句子与句子之间的上下文关联信息。而循环神经网络主要是捕捉序列之间的特征信息，因此根据我们的对话数据特点，本文设计并构建了一个基于双层循环神经网络的匹配模型。对于循环神经网络结构，在不断克服其缺点的道路上，衍生了很多变体，主要是对其循环单元提出了不同的计算方式。比较著名的有长短期记忆网络(LSTM)和门限循环单元(GRU)。相较于长短期记忆网络，GRU 具有与之相当的实验效果，同时运算参数更加简洁，计算速度更快。经过简单的实验验证，我们最终选择了 GRU

作为模型的基本循环计算单元。由于句子的词与词之间和同一组对话的句子与句子之间均存在着丰富的场景信息，因此我们构建了层级的循环神经网络分别对每个句子和句子组进行建模。模型的基本结构图如图 4-3 所示，q 表示问句，a_*表示答句，当然 q 也可表示答句，此时对应的a_*则表示问句，w_*表示对应句子中的字或词向量，h_*表示计算单元的输出状态。模型主要包括三个方面：基于 GRU 的句子建模，基于 attention 的信息自动归纳以及答案置信度排序与阈值选择。以下就这三方面作详细阐述。

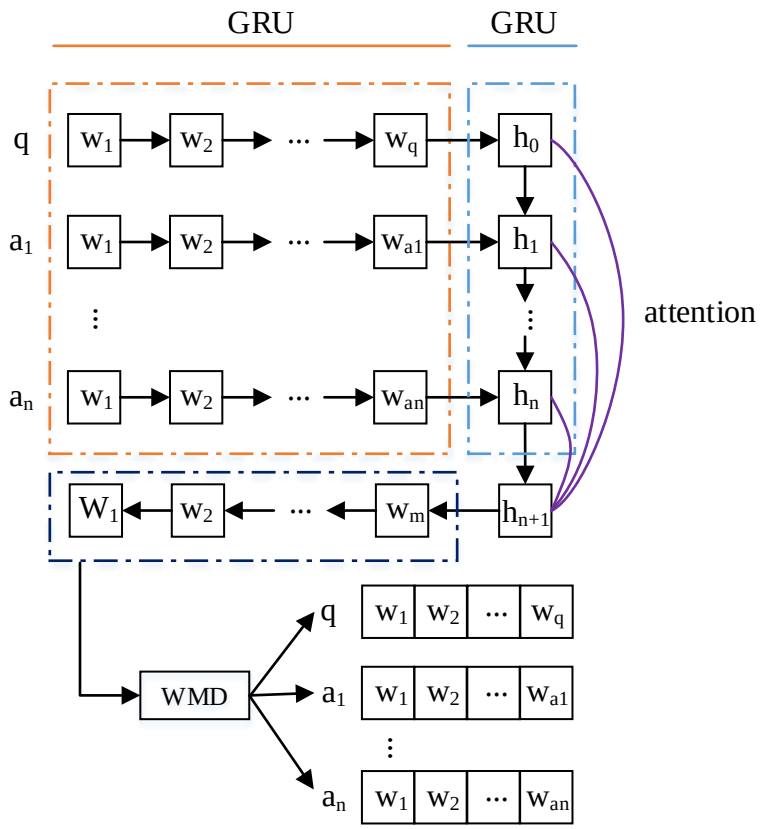


图 4-3 基于 RNN 的问答匹配关系识别模型

4.3.1 基于 RNN 模型的语句向量表示

门限循环单元(GRU)是由 Cho 等人[42]于 2014 年提出的，它主要是在 LSTM 的基础上，将遗忘门和输入门合成了单一的更新门，同时混合了细胞状态和隐藏状态，因此参数更加简单，是现在非常流行的循环单元变体之一。具体的单元结构如图 4-4 所示。

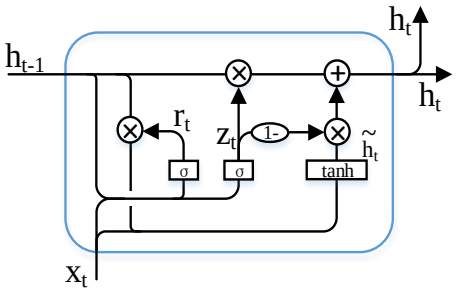


图 4-4 GRU 单元结构图[2]

具体计算方式如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{4-1}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{4-2}$$

$$h_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \tag{4-3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \tag{4-4}$$

其中，z_t表示更新门，r_t表示重置门，h_{t-1}表示上一状态，x_t表示当前输入状态，h_t表示当前输出状态。从公式(4-3)和公式(4-4)我们可以看出，重置门决定上一隐藏状态的比重，更新门决定新的隐藏状态所占的比重，这里使用的激活函数为 tanh。

首先，我们会通过词向量将句子转换为一个二维矩阵，由于使用的是 batch 训练，因此输入 GRU 句子建模的矩阵实际是三维的，大小为 max_len×batch×dim。max_len 表示句子的最大长度，batch 表示一个训练批次的句子个数，dim 表示词向量的维度，这里我们选择 200 维。如图 4-5 所示，经过第一层的 GRU 句子建模后，输出了一个大小为 batch×dim 的句子特征矩阵。

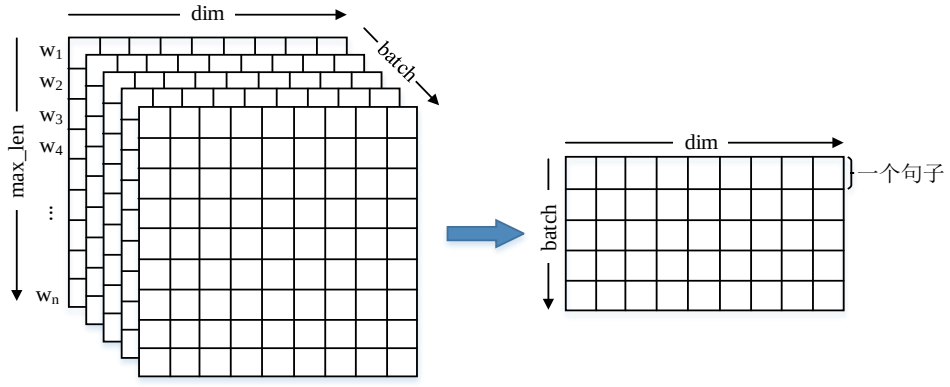


图 4-5 GRU 句子建模

（1）使用 batch 进行训练

[2] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

选择一个合适的批次大小对于模型的训练很重要。我们的训练是基于 batch 计算的，其主要是针对非凸损失函数和数据集大小而来。批表示了所有数据样本上的抽样表现，为了防止达到局部最优，这种方式相当于引入了修正梯度上的采样噪声。批次大小的选择，某种程度上决定了梯度下降的方向。当数据集过小时，我们可以考虑将 batch 扩大到全局数据，结合适当的优化算法完成梯度更新。当数据较大时，batch 太大内存容量无法满足需求，计算时间长，参数修正缓慢。batch 太小，内存利用率低，训练震荡大，难以收敛。在批次大小和训练效果之间找到一个平衡，很重要，提高内存利用率的同时，确定了更准确的梯度下降方向，收敛更快，减少了训练时间。

（2）补零计算的处理

在利用 GRU 对句子建模时，保留了句子中所有的词信息，以句子的最大长度作为顶层神经网络句子输入矩阵的第一维长度。句子的长度不一致，导致我们无法获得维度大小一致的输入矩阵。为了满足这一要求，我们对长度不够的句子矩阵做了补“0”的处理。但这样就会带来一个问题，如果我们按照常规的 RNN 计算，就会因为零向量的填充，导致结果错误。针对这一问题，我们对句长做了标记，当遇到非句子中的词向量时，回滚上一次的计算结果，保证了填充零向量的过滤，不对计算结果产生影响。

4.3.2 基于 attention 的信息自动归纳

从基本的循环神经网络，到长短期记忆网络、门限循环单元，都是在致力于模型能够从信息序列中自动选择有效的特征信息，但当循环序列过长，模型可能就会对早期记忆的东西遗忘，即使是针对长距离依赖学习的 LSTM、GRU，它们仍然是将所有信息编码成固定长度的向量，在一定程度上也存在这种情况。注意力机制允许网络返回到输入序列，让网络模型再次访问编码器的隐藏状态，从记忆中找回遗失的特征信息。本质上，就是对隐藏状态的加权组合，这样可以通过反向传播算法，动态的加强需要的特征信息。

主流的计算公式如下，m_t表示目标状态，m_s表示输入序列：

$$a_t = \frac{\exp(f(m_t, m_s))}{\sum_{s'} \exp(f(m_t, m_s))} \tag{4-5}$$

$$f(m_t, m_s) = \begin{cases} m_t^T m_s \\ m_t^T W_a m_s \\ W_a[m_t : m_s] \\ v_a^T \tanh(W_a m_t + U_a m_s) \end{cases} \tag{4-6}$$

考虑到对话数据中，同一组句子间存在着强关联的上下文信息，具有一定的场景特点，如图 4-3 所示，本文主要在句子之间的编码部分设置了 attention。具体计算过程是：设计一个函数，将目标状态与输入状态通过非线性运算联系起来，然后通过送入 softmax 函数做归一化得到概率分布，作用于输入状态，经非线性运算添加入目标状态。这里的目标状态指经第二层循环神经网络编码后的最终状态。

4.3.3 句子置信度排序与分类阈值选择

经过上一节的第二层编码后，获得了每个句子组编码的最终状态。通过该最终状态，生成与问题对应的答案或者与答案对应的问题，其实质是一个解码器，最初应用于机器翻译，后来推广到文摘、问答等任务上。基本结构如图 4-6 所示。

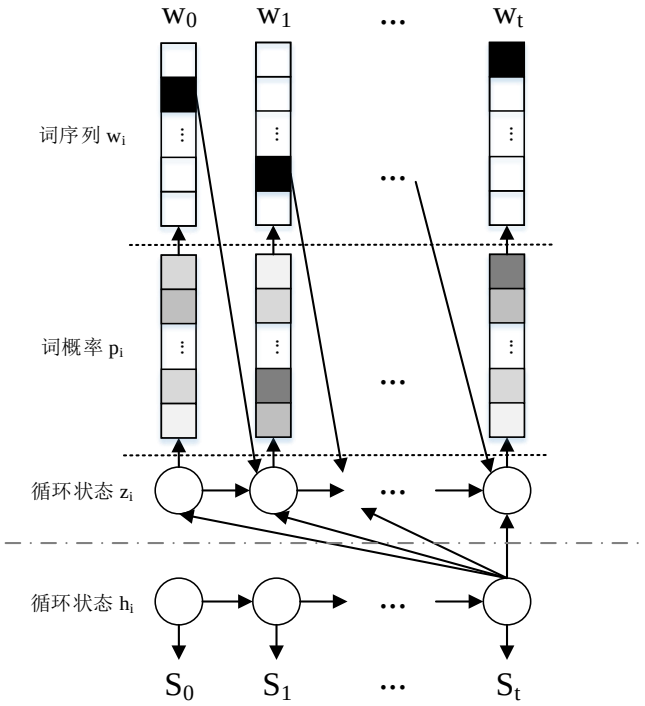


图 4-6 答案生成解码图

以下以生成答案为例作说明。在模型的训练中，以一轮对话中完整的答案作为类标，交叉熵为代价函数，不断计算目标句子与生成句子之间的交叉熵。选择由 Ilya Sutskever 等人[48]提出的 Nesterov Momentum 作为优化函数。针对 SGD 存在更新方向完全依赖于当前 batch，更新十分不稳定的情况，动量(Momentum)被引入。它在更新的时候在一定程度上保留了之前的更新方向，同时根据 batch 微调更新方向，具有一定的稳定性，收敛更快。Nesterov Momentum

是在 Momentum 的基础上的一种改进，每次在 Momentum 的更新方向上继续更新一步，经实验验证，学习速度更快。

生成答案后，在答案与对应问题的 5 条下文之间进行词移动距离的计算，由此，便可得出答案的置信度排序，同时根据阈值的设定对下文进行答案的分类。由于 GRU 生成模型设计的是以 6 条语句一组，也就是一个问题和五条答案一组，进行答案的生成。在进行社区问答数据的实验时，其受限于每组 10 条答案下多数是有大于 1 条的正确答案。与之对应的，会将测试数据的每一问答组分为两组，最终原始问答组中答案的置信度取两小组的生成答案与原始问答组中答案计算的距离之和。选择词移动距离(WMD)原因主要有两点：首先，相较于其他距离计算方法，WMD 在很多任务的语义距离计算上，具有明显的优势，这点在其发表的论文中也得到了验证。再则，词移动距离的计算不受语序和重复词的影响，这对于我们生成的信息没有很大限制，只要是包含与准确答案接近的信息即可，而对表达形式对于置信度的计算不产生影响。

4.4 实验环境

本课题的实验主要包括两大部分：实验数据的统计分析、预处理和抽取工作，基于深度学习的问答匹配关系识别和语句补充关系识别的模型训练和结果测试。因此，主要硬件资源包括一台 PC 和一台服务器。软件环境主要包括 python2.7，theano，以及一些自然语言处理、矩阵运算和统计机器学习方法相关的库，如 jieba，scikit-learn，numpy 等等。具体的硬件配置环境如表 4-1 所示：

表 4-1 硬件资源环境配置

设备	资源名称	配置
1 台 PC	操作系统	Windows8
	CPU	Intel core i5
	内存	16G
	硬盘	500G
1 台服务器	操作系统	Ubantu12.04
	CPU	2*10 核
	GPU	2 块 Tesla K40m
	内存	128G

4.5 答案排序与选择的评价方法选择

一般的，在人工对话的生成中，会通过人工检测的方式对结果进行评判，

但这样需要较高的人力成本和时间成本，无法实时的对结果进行评价。这里，我们采用一种排序的方式对生成的结果进行评价，生成的信息有可能并不是完整的，无法主观的对其进行评价，我们通过生成的信息对原本组别中的句子进行排序的方式，测试生成信息的可靠度。同时，对已打好分的句子，给定阈值进行正负样例的分类，给出分类的评测。由此，便可抽象出本课题的两类主要评价指标，分别为排序和分类，以下就本文使用到的这两类指标进行详细介绍。

4.5.1 答案排序评价方法

在排序指标中，我们主要使用了 MRR(Mean Reciprocal Rank)和 MAP(Mean Average Precision)两个评价指标，这两个指标是 TREC QA 和 Semeval QA Track 中常用的两种指标。计算公式如下：

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{4-7}$$

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{4-8}$$

$$AveP(q) = \int_0^1 p(r)dr$$

式中 Q——表示 query 集合；

$rank_i$——表示与 query 最相关的结果的排序；

$AveP(q)$——表示平均精度，是一个 q 下多个相关文档的排序的倒数之和求平均。

MRR 衡量的评估假设是基于唯一的一个相关结果而来，也就说针对每个 query 只评估一个最相关的结果。MAP 反应的是全部相关文档上的性能的单值指标，衡量的是每个 query 下所有相关文档的排序。

对应到本文数据中的评价，当对话数据中的上文或者下文中，相关语句中有任一个排名越靠前，MRR 就会越高；当所有相关语句的排名整体越靠前，MAP 就会越高，这正符合了我们对测试结果的评价要求。

4.5.2 答案选择评价方法

在分类任务中，比较常用的指标就是准确率(Precision，P)，召回率(Recall，R)和 F1 值。准确率和召回率是广泛应用于信息检索和统计分类领域的两个重要指标。准确率又称查准率，主要用来衡量识别出的结果中正确的比例。召回率又称查全率，主要用来衡量识别正确的结果在正确的结果中的比例。计算公

式如下：

$$P = \frac{TP}{TP + FP} \tag{4-9}$$

$$R = \frac{TP}{TP + FN} \tag{4-10}$$

式中 TP——表示正样本被判定为正样本的个数；

FP——表示负样本被判定为正样本的个数；

FN——表示正样本被判定为负样本的个数。

准确率和召回率之间均有一定的制约性，我们无法单独用其中一个指标去衡量实验方法的性能。需要在准确率和召回率之间取得一个平衡，综合考量两者的结果。这里使用 F 值对两者进行综合评价，计算方法如公式(4-11)所示：

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{4-11}$$

其中，参数 β 表示 P 和 R 的权重等级，这里我们取 β=1，表示准确率和召回率具有相同的权重，也就是我们所说的 F1 值。

4.6 实验结果对比与分析

本章节实验的设计主要针对模型在英文社区问答数据集和中文对话数据集上的表现进行对比，针对不同模型在同一类型的数据集上的效果进行对比，以及在对话数据上不同抽取形式的数据进行对比分析。

4.6.1 cQA 数据的实验结果对比分析

本实验使用 Semeval-2016 社区问答数据的验证集进行训练，测试集进行测试。没有使用训练集的原因在于在验证集上训练测试 CNN 模型效果与在训练集上训练测试的模型效果比偏低，但并不是相聚甚远，这里主要参照的是本实验室在 semeval-2016 上参赛初次提交的官方结果。而在本次的实验中，主要是当 CNN 模型的测试结果达到一个合理的范围后，在同样的数据下对比基于 GRU 的生成模型与基于 CNN 的模型效果。本组数据下的实验，选择了一个更大的语料-维基百科英文语料训练词向量作为实验模型的基本词向量。实验主要分为两部分，一是在基于 GRU 的生成模型上进行了不同模型结构与不同数据标签的对比，二是在不同种类的模型上进行了实验对比。

首先，在基于 GRU 的模型上进行了不同方式的对比，其对比结果如表 4-2 所示。

（1）分别将生成模型的标签数据设为问题和答案

其实际意义就是在一组包含一个问题和多个答案的数据生成答案的时候，分别是参照问题的信息和答案的信息得来的。主要目的是观察在社区问答中，问题和答案对于生成答案而言，两者的信息贡献度。从表 4-2 的一二行数据中，能够明显的发现，答案的贡献度显然高于问题的，在生成模型上的实验，我们以答案作为生成信息的参照。

（2）添加 attention 机制

这里主要是对组内问题和答案经过第一层编码后进行 attention，也就是做不同句子级的注意力机制。通过以下对比结果，我们可以发现，在进行句子编码信息不同权重的添加后，结果反而有所降低。在社区问答中，由于并不存在强上下文关联信息，更注重的是问句和答案之间的直接匹配关系，这里对组内所有句子信息进行不同权重的引入，有可能带来的噪音大于正确信息的加强。

表 4-2 基于 GRU 的不同方式的实验在 cQA 数据上的对比结果

实验方法	label	MRR	MAP
hierarchy GRU	Question	0.61120	0.55799
hierarchy GRU	Answer	**0.66726**	**0.58914**
hierarchy GRU + attention	Answer	0.63776	0.57907

然后，在不同方法上进行了对比，包括两大类深度模型基于卷积神经网络和基于 GRU 的层次模型。值得注意的是，进行 CNN 模型的训练时，我们使用的类标为(0,1)，表示两个句子是否匹配，显然，将答案选择问题转换为了语句的语义匹配问题。但在进行 RNN 模型的训练时，类标却是一个答案句子。从类标，就明显看出了两种模型的内在原理与实现思路上是完全不同的。基于 CNN 的模型是经过句子并行建模后，对拼接特征进行了映射和分类。而基于 RNN 的模型是一种生成模型，主要是对问答组的信息进行了归纳，生成一个语句，通过该语句与正确答案之间的置信度进行排序。由此，这也是我们进行这两种模型对比的初衷，观察基于匹配和基于生成的模型在社区问答答案选择上的表现效果。对比结果如表 4-3 所示：

表 4-3 不同实验方法在 cQA 数据上的对比结果

实验方法	label	MRR	MAP
Baseline 1(random)[49]	(0,1)	0.5871	0.5280
Baseline 2(chronological)[49]	(0,1)	0.6783	0.5953
CNN	(0,1)	**0.77768**	**0.69022**
hierarchy GRU	Answer	0.66726	0.58914

这里的基准方法来自于文献[14]中，Semeval-2016 问答评测专项提供的测

试结果。从与基准方法的对比上，我们模型还是具有一定的有效性。在此基础上，从 CNN 匹配模型与 RNN 生成模型的对比来看，CNN 的效果明显要高，这里的 CNN 匹配模型只注重两个句子之间的语义匹配关系，而基于 RNN 的生成模型关注的是整个句子组的序列特征，其中存在大量的噪音。这也进一步证明了在具有弱场景信息和多条数据的数据组上，直接使用 CNN 的模型匹配效果要优于 RNN 的生成模型的效果，RNN 模型在噪声较多的情况下并不利于信息的自动筛选。

4.6.2 对话数据的实验结果对比分析

上一小节主要做了不同模型在社区问答数据上的对比实验效果分析，这一小节主要做不同模型在对话数据上的对比实验分析，原因在于社区问答数据与人工客服对话数据有较大的不同点，尤其是在交互的场景信息上。以比较不同类型的模型在不同类型的数据上的差异以及交互式问答数据上适用的模型类型和原因。

对话数据的实验上，在与 3.2 节对应的两种思路对应的方式抽取出的两种实验数据上，分别进行了基于 CNN 模型和基于 RNN 模型的对比实验，以及基于 GRU 的不同方式的对比实验，最终对两种思路的实验结果进行了对比以及优劣性的分析。

策略一 以每轮对话为基础分析

每轮对话为基础是指以一轮对话为一组，每组的第一句是该轮对话中语义角色为客户的语句，可能是完整的问句也可能是不完整的问句，该组中紧接着有五句下文,其中包含该轮对话中问句的完整答案,可能是一句也可能是多句。针对此种类型数据，进行基于 GRU 的模型和基于 CNN 的模型对比实验，实验结果如表 4-5 所示；同时针对基于 GRU 的生成模型做不同类型的对比实验，实验结果如表 4-4 所示。其中，实验方法以模型中种类和数据中问题的完整性区分，hierarchy GRU 表示层级 GRU 模型，complete q 表示数据的每个问答组中第一条问题语句是完整的，相反，single q 则表示只选取了具有补充关系的问题中的一句，问句是不完整的，该轮对话中的问题不存在补充关系除外。

（1）基于 GRU 不同方式的对比实验

首先，将 GRU 生成模型的类标分别处理为每论对话中的完整问题和完整答案进行对比实验，意义在于探究对话数据中问题与答案之间简单的语义关联程度以及两者对于对话组生成答案的信息贡献度。从表 4-4 的一二行，可以看出，对话数据上答案对信息归纳的贡献度要优于问题对信息的贡献度，同时，对话数据中问题和答案之间的语义相关性特征更高级，并不是简单的特征与距

离计算就可以完成的。

再则，将完整的问题与不完整的问题对实验的影响进行对比，主要是为了探究在对话数据上，问题的完整性对于对话中场景信息的影响以及由此造成的对于答案生成的影响。从表 4-4 的二三行结果可知，从排序的角度上评价，单个不完整的问题与完整的问题在实验结果上是有一定差距的，问题的完整性影响了对话场景信息的完整性，而基于 GRU 的生成模型主要依赖对话中的上下文特征进行信息归纳的，对话中问题的不完整影响了最终生成模型根据场景信息进行的信息归纳，使用问句完整的数据训练模型对实验效果有一定的提升。

表 4-4 基于 GRU 的不同方式的实验在对话数据一上的对比结果

实验方法	label	MRR	MAP
hierarchy GRU/complete q	Question	0.53999	0.52768
hierarchy GRU/single q	Answer	0.82024	0.79827
hierarchy GRU/complete q	Answer	**0.86986**	**0.84409**

（2）不同类型的模型方法的对比实验

不同模型主要包括基于 CNN 的匹配模型、基于层次 GRU 的生成模型以及添加注意力机制的层次 GRU 生成模型。从表 4-5 中类标的说明便可看出，一类是匹配排序模型，一类是生成再排序的模型，主要目的是根据两种模型的原理不同，探究对话数据中上下文信息的强弱，以及模型对数据中场景信息的特征抽取能力。以下的对比实验均是基于具有完整问题的对话组而完成的。

从表 4-5 可以发现，基于 GRU 的循环神经网络的效果明显优于基于 CNN 的匹配模型的效果。这不仅说明，对话数据中存在丰富的场景信息，同时基于 GRU 的层次循环神经网络对于这种信息的特征抽象要比基于并行 CNN 的简单匹配要优越得多，在基于 CNN 的匹配中，并未考虑到其他句子的特征信息，而在基于 GRU 的模型中，是对一组问答句进行建模和特征学习，而 GRU 的循环神经网络就是对序列的一种学习，在循环处理词向量矩阵的过程中，保证了对词与词之间和句子与句子之间关联信息的特征抽取，导致了相比于并行 CNN，层次 RNN 学习场景信息更优越。但是，当在层次 RNN 模型上加入句子级的 attention 机制后，效果却有所下降。3.3 节的数据分析表明，本对话数据中有接近百分之 60%的一对一匹配，也就说在一组数据中很多是下文中只有一个答案句与问题匹配，这就表示对话组中存在大量的噪音数据，同时这些噪音数据有可能存在是同一主题而非同一轮对话的语句，这就导致了模型的训练结果可能会向答案以外的句子偏离。因此，在这里的对话数据中反而是基于 GRU 的层次模型效果比较优越。

表 4-5 不同实验方法在对话数据一上的对比结果

实验方法	label	MRR	MAP
CNN/complete q	(0,1)	0.75978	0.74808
hierarchy GRU/complete q	Answer	**0.86986**	**0.84409**
hierarchy GRU + attention/complete q	Answer	0.82145	0.79812

但是，以上表格中的结果是对正确答句的排序评价结果，我们希望模型不仅能排序正确，同时通过阈值的阈值的选择，具有比较稳定的分类能力。

因此，在基于 GRU 的两种模型的置信度结果上，通过设置不同的阈值，测试了对话数据中答案分类的 F1 值，主要目的在于测试经过这种阈值的选择是否具有稳定性，同时是否能达到一个比较好的效果。从图 4-7 答案置信度选择的测试中发现，在基于 GRU 的层次 RNN 模型上，阈值区间在[0.25,0.5]时，分类效果比较稳定优越；在加入了 attention 的层次 RNN 模型上，阈值区间在[0.2,0.5]时，分类结果比较稳定优越。由此可知，这种基于 GRU 的层次模型，通过阈值的选择进行分类效果还是不错的，同时阈值在一个比较稳定的区间上，对于阈值的确定会比较容易一些。不过，可以发现，加入了 attention 之后，虽然在排序结果上有所降低，但在阈值的选择范围上却有所提高。

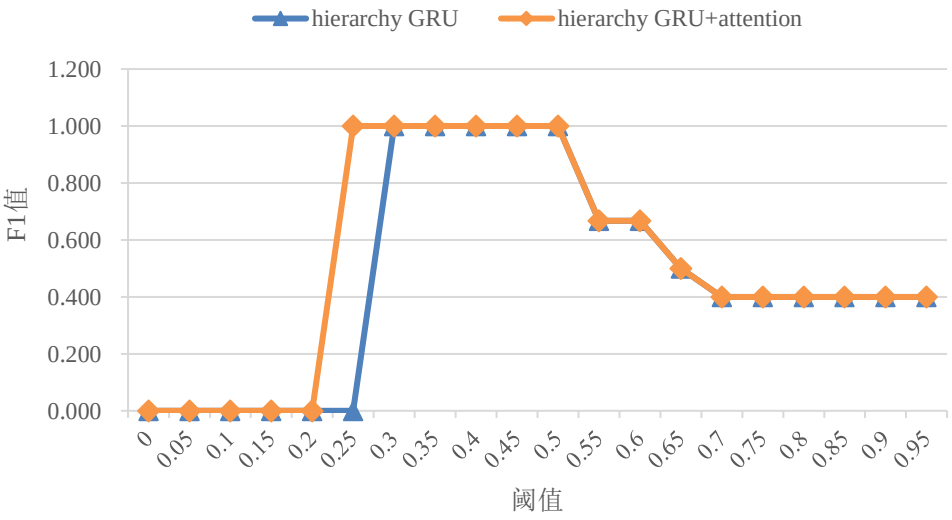


图 4-7 策略一基于 GRU 的答案置信度选择与分类结果

策略二 以滑动窗口抽取对话分析

以滑动窗口抽取对话是指以语义角色为客服的语句为基础，以固定窗口大小滑动，抽取语句组。这里的语句组包含一条客服所说的基准语句和该语句的多条上文，实验中，我们取窗口大小为 6。在抽取出的每组对话中，上文将会包

含与答句匹配的问句，与答句不匹配的问句以及与答句具有补充关系的答句和与答句不具补充关系的其他轮对话中的答句。这里的问句指语义角色为客户的语句，答句指语义角色为客服的语句。在此种对话数据上，分两步进行实验。

（1）基于 GRU 的不同方式的对比实验

首先，进行添加注意力机制的对比。从表 4-6 的二三行，可以发现，当加入了句子级的注意力机制后，结果反而有所降低，这里的 attention 是对该组内所有句子编码的 attention，当相似度较高的句子噪音存在时，反而会导致结果下降，比如非同一轮但意图相近的语句，其主要来源于在同一主题下的不同轮的连续对话。

再则，是编码方向的影响。保持第一层的正向编码方向不变，分别调整第二层为正向编码和反向编码。从表 4-6 的三四行，可以发现第二层反向编码的效果更好，这表明了在对话数据中，进行场景信息归纳时，其更依赖于短程信息，将需要归纳的信息放置于最终状态越近的位置，越有利于归纳实验效果的提升。

（2）基于不同类型的模型的对比实验

这里的不同模型包括基于 CNN 的匹配排序模型和基于 GRU 的层次模型。从表 4-6 的结果来看，在基于滑动窗口抽取的对话数据上，对话场景信息未必完整的情况下，反而基于 CNN 的简单匹配效果更好。在场景信息不完整情况下，层次循环神经的对信息的特征提取能力具有一定的劣势，此时直接对问答对进行并行的特征映射进行全连接层的特征非线性运算优越于依次将信息进行循环编码。

表 4-6 不同实验方法在对话数据二上的对比结果

实验方法	Label	MRR	MAP
CNN	(0,1)	**0.78651**	**0.77441**
hierarchy GRU	Question	0.71338	0.69268
hierarchy GRU+attention	Question	0.62916	0.61523
forward_backward GRU	Question	0.74880	0.72538

以上分析的均是策略二数据的匹配排序指标结果，图 4-8 展示了对排序的置信度进行阈值的选择和分类的结果。从图中可以看出，在策略二抽取出的数据上，这种简单基于层次 GRU 的模型并不能通过选择阈值进行很好的分类。加入 attention 后，会在某个阈值点达到一个峰值，不具有稳定的阈值选择范围。相较于其他两种 GRU 模型，正反向 GRU 模型具有更稳定的阈值选择范围和更好的分类效果。但与策略一的阈值选择的稳定性和分类效果相比均具有一定的

差距。

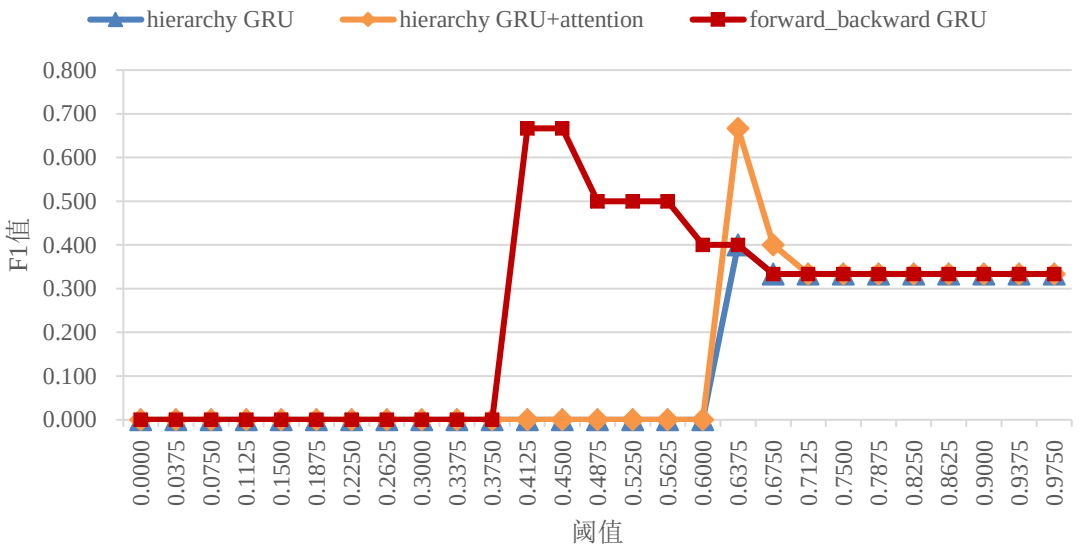


图 4-8 策略二基于 GRU 的答案置信度选择与分类结果

4.6.3 策略一和策略二的实验结果对比分析

从表 4-7 中，可以看出，在排序指标的评价上，策略一的抽取方式要明显优于策略二的抽取方式。在对话中具有完整场景的情况下，更倾向于使用基于循环神经网络的模型对序列信息进行建模和特征抽取，在对话不完整的情况下，基于 CNN 简单句对匹配或者其他基于句对进行特征提取的深度模型可能会具有更好的效果。从 4.6.2 节的阈值选择和分类结果上，可以观察到策略一也具有一定的优势，不仅在阈值选择的范围上比较大，具有一定的稳定性，这对于我们进行阈值选择时，就不会太严苛，同时分类的 F 值也会比较高。

表 4-7 策略一和策略二的实验结果对比

实验方法		MRR	MAP
策略一	CNN	0.75978	0.74808
	GRU	**0.86986**	**0.84409**
策略二	CNN	0.78651	0.77441
	GRU	0.74880	0.72538

4.7 本章小结

本章主要介绍了人工客服对话数据中，本文使用的方法的实现细节，包括基于 CNN 的匹配排序模型和基于 GRU 的层次循环神经网络的生成模型。在基

于 GRU 的模型上，分别对两层编码过程、句子之间的注意力机制、解码过程和答案的置信度计算和选择、答案分类进行了详细介绍。接着，简单描述了实验环境。最终，对不同抽取策略上不同模型的对比实验细节进行了介绍和结果分析，对不同抽取策略的实验结果进行了分析，基于 GRU 的层次循环神经网络在完整会话轮上的问答匹配具有一定的优越性。

第 5 章 语句补充关系的识别

5.1 引言

在前一章问句完整性的对比实验上，本文发现在问句完整的情况下，基于完整会话轮数据的基础上实验效果更好，这表明问句的完整性即问句补充关系的识别是交互式问答中的一个关键问题。同时在上一章节的实验中，问句的补充关系是由人工标注而来，在实际的使用中，我们需要通过机器学习方法来学习这种关系标注以代替人工，由此衍生出了交互问答中的第二种语句关系研究，语句补充关系的识别。

本章主要介绍语句补充关系识别的设计思路和解决方案，主要将语句之间的补充关系识别转化为了补充关系的二分类方法进行解决，包括基于支持向量机的补充关系分类、基于卷积神经网络的补充关系分类、基于循环神经网络的补充关系分类。详细阐述了这些方法在补充关系识别中的使用，并描述了数据的构建方法，给出模型结果的分析。

5.2 语句补充关系识别的方法介绍

本文将语句补充关系的识别，转化为了多个语句之间匹配关系的分类，由于这种补充关系并不是普通意义上的语义匹配，与之不同的是，他们之间不仅可能有简单语义的匹配，还可能存在意图、主题上的相似性，互相之间的一种补充关系，我们不能用普通的文本相似度计算方法来完成。因此，我们选择了三种分类方法来进行补充关系识别的比较，分别是传统的支持向量机模型、基于卷积神经网络的关系分类模型、基于循环神经网络的关系分类模型。

5.2.1 基于 SVM 的语句补充关系识别

相较于其他传统统计机器学习算法，支持向量机性能稳定，求解是一个全局最优结果，不同于部分机器学习算法贪心选择的局部最优，对于小样本泛化能力比较强，在非线性分类及高维模式识别中表现出许多特有的优势。由于本文的样本数据有限，因此选择了支持向量机作为对比实验中的浅层机器学习方法。以往的许多机器学习方法都是求经验风险最小化，但真实风险未必是最小化，经验风险未必能够逼近真实风险，这就导致了在样本集上的测试结果极高，但在真实分类时效果极差。不同的是，SVM 是基于结构风险最小化理论，在模

型的复杂性和学习能力之间寻求最佳折中，得益于其核函数和松弛变量，支持向量机具有较强的泛化能力。

使用支持向量机主要在于特征的抽取和核函数的选择。本章节所完成的任务实际上是两个句子之间的补充关系的判定，模型的输入是两个句子的特征表示，模型的输出是两个句子补充关系的判定。其中，使用的特征，是两个句子词向量的平均表示拼接而成，然后经线性核函数映射到高维特征空间，确定最大分隔超平面和支持向量，由决策函数给出线性分类结果。这里使用的是线性核函数，与高斯径向基核函数 RBF 相比，效果略优。

5.2.2 基于卷积神经网络的语句补充关系识别

这里使用的卷积神经网络结构，与 4.2.1 所介绍的结构相类似，见图 4-1 所示。同样是将两个句子通过卷积神经网络并行卷积拼接后送入多层感知器进行分类，不同之处在于我们这里的多层感器的网络深度不同，输出的结果也有所不同，如图 5-1 所示。由于本章节的任务是分类问题，在网络的最终输出会修改为类标签而非置信度。

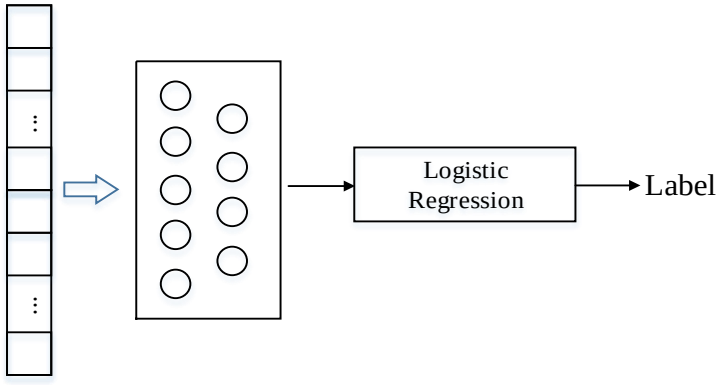


图 5-1 基于 CNN 的问句补充关系识别模型 MLP 及分类器基本结构

具体的网路节点参数如下：

经统计，句子补充关系识别的数据中，90%以上的句子长度达到 30，句子平均长度大约在 25，因此，我们选择了输入矩阵大小为 30*200 维。

特征提取阶段，我们主要使用了三层卷积和三层池化。第一层卷积池化窗口的大小为：filter_shape0=(3,1)，pool_size0=(2,1)；第二层卷积池化的窗口大小为：filter_shape1=(3,1)，pool_size1=(2,1)；第三层卷积池化的窗口大小为：filter_shape2=(3,1)，pool_size3=(4,1)。同时对两个句子使用的相同的滤波器和池化窗口进行特征提取，输出向量的维度为 200 维，最终拼接得到 400 维的向量。

多层感知器选用全连接网络，本章节实验中尝试了一层感知器和两层感知器的实验，第一层隐藏层的输出维度为 100 维，第二层隐藏层的输出维度为 50

维。最终送入 softmax 做分类处理。

5.2.3 基于循环神经网络的语句补充关系识别

循环神经网络主要处理的是具有时序状态的信息，这里选择比较著名的长短期记忆网络作为问句补充关系识别模型的基本神经单元，用意在于捕捉句子内词与词之间的语义关联信息和两句子之间的上下文信息。与以往的循环神经单元不同，长短期记忆网络不仅能捕捉到短程依赖信息，同时也是针对捕捉长程依赖信息提出的一种网络。其基本的单元结构如图 5-2 所示：

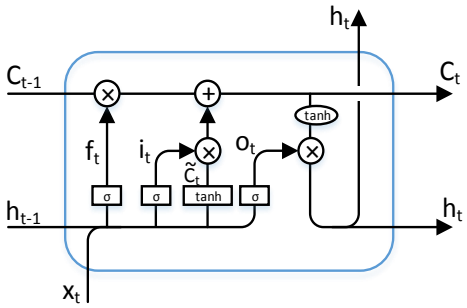


图 5-2 LSTM 单元结构图[3]

LSTM 网络单元主要通过三个门控制信息的信息状态，包括遗忘门，输入门和输出门。f_t表示遗忘门，i_t表示输入门，o_t表示输出门。算法中的计算公式如下所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5-1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5-2}$$

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{5-3}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{5-4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5-5}$$

$$h_t = o_t * \tanh(C_t) \tag{5-6}$$

这里的σ表示 sigmoid 单元。遗忘门主要控制是否丢弃上一单元的信息，在我们的模型中表示以往词向量计算的历史信息。输入门决定更新哪些信息，tanh 层生成一个新的候选值，两者一起生成新的更新状态，以备下一次遗忘门选择信息。输出门决定新的更新状态的输出生成新的隐藏状态，即隐藏层的输出状态。不难看出，LSTM 通过不同门的控制，对句子中词向量的特征信息进行保留或者遗忘。

[3] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

基于这种计算方式的 LSTM 单元，我们使用串联的 LSTM 进行补充关系识别实验。其模型结构如图 5-3 所示：

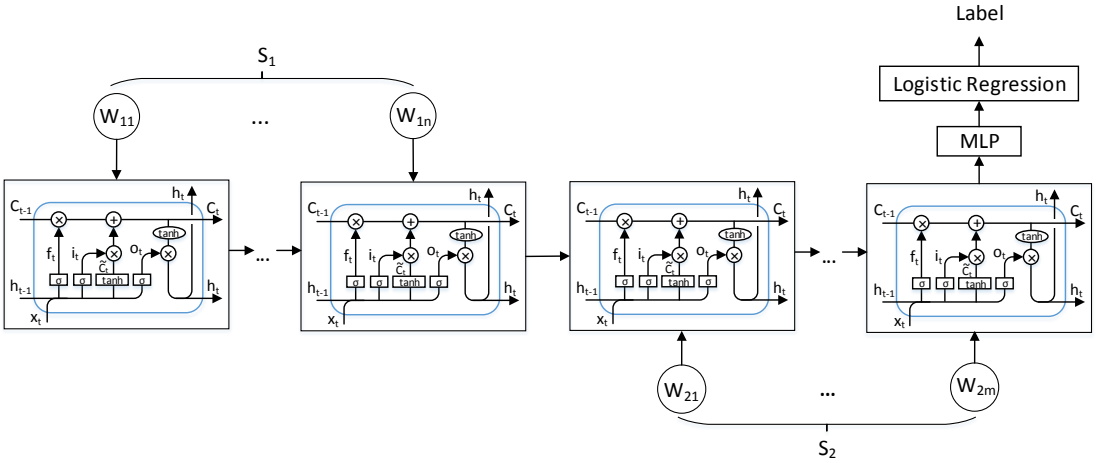


图 5-3 串联 LSTM 问句补充关系识别模型图

输入模型的是一个句对，输出模型的是分类标签。使用词向量分别将句子转换为 n*dim 维的矩阵，进行拼接，送入循环神经网络利用 LSTM 进行特征信息的选择计算，将最终的隐藏状态作为输出状态。送入多层感知器进行特征映射，经非线性运算后，进行逻辑回归分类。

5.3 语句补充关系识别的数据构建

（1）数据构建

针对语句补充关系的识别，本文主要构建了三种实验数据用于实验对比分析。

第一种实验数据，在每组对话中，构建具有补充关系的问题对正例和负例。具有补充关系的问题对来源于该组对话中的每轮对话，是问题对的正例。与之对应的负例来源于对应问题的上文中的问题，可能在同一轮对话中，也可能在不同轮次的对话中，但在同一组对话中。

第二种实验数据，对以上抽取出的数据做过滤处理，在电商客服交流的过程中，存在大量的无意义对话，主要通过规则过滤其中只存在表情，数字以及无意义的客套用语，以及长度太短语句，比如：‘/:809’，‘嘻嘻’，‘嗯，好的’。

第三种实验数据，由于数据量有限，在此基础上我们将具有补充关系的语句扩充到了答句，即将具有补充关系的问题和答句均作为我们的实验数据，改变了负例构造的方法。从上一组对话中随机选择一句与当前问题或者答句组成语句对负例。同时过滤掉部分噪音语句，主要包括短句、部分无意义语句，过滤方式同第二种实验数据。

（2）词向量说明

本次任务中，主要训练了两类词向量，一类是基于本次实验数据的词向量，一类是基于以上同于领域语料的字向量。使用这两类原因在于对比在词和字以及基于不同语料的词向量下，实验结果的差异。具体方法与参数已在 3.5 节阐述，这里不再赘述。

5.4 语句补充关系的对比实验结果

本章节的实验设计主要针对上一节中提到的三种实验数据进行，分别为未经过滤的问句补充关系识别，经过滤的问句补充关系识别以及不同负例构造的问句或者答句的补充关系识别。其中使用了三种类型的实验方法进行了对比，包括支持向量机的方法，基于 CNN 的匹配模型和基于 LSTM 的匹配模型。

实验一 问句补充关系识别

（1）未经过滤的问句补充关系识别

本实验中语句补充关系的识别针对的是对话数据中语义角色为客户的问句，包含三大类和十五个小类的主题类型，包括礼貌用语和闲聊部分的语句，负例的构造来源于同一组对话数据中该问句的上文。使用的词向量为对话数据训练所得的词向量，共计 2748 词。训练集 5296 条，测试集 590 条。

实验中主要使用了三种模型方法，包括支持向量机，基于并行 CNN 的匹配模型和基于串联 LSTM 的分类模型。从表 5-2 可以看出，基于 SVM 的模型效果最优，但基于串联 LSTM 的方法效果不是很好。本数据中包含的信息种类比较多，尤其是包含类型为“闲谈”和“礼貌用语”的语句类型或者表回应的无实际意义的语句，语句对比较难以判别。

表 5-1 未过滤数据噪音样例

组别	内容		补充关系
1	好的	我自己备注可以不	0
2	好的	我拍了就差付款	1
3	只要到就好	等太久啦	0
4	只要到就好	/:815	1

比如样例表 5-1 中，每组两个句子，表示是否有补充关系。“0”表示非补充关系，“1”补充关系。“好的”事实上只是客户的一个简单回应，并没有包含实词的含义，属于百搭型，但在这里与不同的语句搭配后，便具有了不同的补充关系，对于机器来说辨别比较困难，此外，类似“/:815”这样的表情符号，使用也比较多，这使得另一组的负例比较难以区分，第三、四组中也许我们会觉得第三组也为正例。此时，模型对于这种类型的自动学习特征非常困难。浅

层模型通过指定特征的分类反而效果也许会更好一些。

表 5-2 未过滤的问句补充关系识别对比结果

实验方法	P	R	F1
SVM	0.547	0.850	0.666
CNN	0.549	0.736	0.629
Series LSTM	0.490	0.237	0.320

（2）已过滤的问句补充关系识别

考虑到上一实验噪音数据对实验效果的影响，我们对上一实验数据进行了过滤，包括类型为“闲聊”和“礼貌用语”的无意义语句，以及通过规则对短句且其中包含类似“好的”,“恩，谢谢”的语句、表情符号、单个数字等无实际意义的语句进行过滤。得到已过滤数据，其中，数据集 3012 条，测试集 356 条，依然使用本客服对话数据训练所得的词向量。

从表 5-3 可以看出，与上一实验相比，效果反而有所下降，理论上过滤之后的语句对应该更利于模型的特征学习。但是，深度学习模型的是应该建立在数据量比较大的基础上,过滤后我们的数据量更小,也许会不利于模型的学习。因此，我们又进行了第二个实验，扩充补充关系识别的范围。

表 5-3 已过滤的问句补充关系识别对比结果

实验方法	P	R	F1
SVM	0.605	0.590	0.597
CNN	0.522	0.719	0.605
Series LSTM	0.507	0.427	0.463

实验二 问句和答句补充关系识别

本实验中，我们扩大了实验数据的范围，在问句补充关系的基础上加入答句的补充关系。在补充关系的语义匹配上，同类型对话中，两者存在一定的共性。并且在构造负例时选择不同组别的语句与当前问句或者答句组合为负例，过滤包含如“好的”、“嗯嗯”的短句，以及单个的表情符，如：“/:>_<”，“/:079”等和单个数字等这样无意义的语句。最终，训练集共计 9744 条，测试集共计 1084 条。为了测试不同词向量的效果，这里我们分别使用了对话数据的词向量和百度通用语料的字向量。

首先，我们使用基于 CNN 的匹配模型进行了对比实验。包括改变多层感知器的深度，改变词向量的种类和规模。从表 5-4 的结果可以看出，扩大词向量的规模，使用通用语料的字向量效果最好。在使用原客服语料词向量的情况下，增加多层感知器，效果也会有提升，但不及基于字的通用原料的提升高。

表 5-4 基于 CNN 模型的对比实验

实验方法	P	R	F1
CNN(tb_w2v_words)	0.655	0.640	0.647
CNN(tb_w2v_words)+more mlp	0.681	0.653	0.667
CNN(ty_w2v_char)	0.633	0.731	0.678

然后，进行了基于串联 LSTM 的不同对比实验，主要是不同词向量之间的对比。从表 5-5 的结果发现，改变词向量为字向量并扩大词向量规模后，实验效果会有所提升，但是仍不及基于 CNN 的匹配模型。这里并没有采用并行建模的方法，而是进行了串联的循环序列建模。

表 5-5 基于串联 LSTM 的对比实验

实验方法	P	R	F1
Series LSTM(tb_w2v_words)	0.500	0.784	0.611
Series LSTM(ty_w2v_char)	0.500	0.867	0.634

最后，我们进行了不同模型的对比实验分析，主要包括支向量机模型，并行 CNN 匹配模型和串联 LSTM 分类模型。在 SVM 中，每个句子分别使用 200 维的特征进行表示，每个句对 400 维的特征表示。在 CNN 匹配模型中，句对并行建模，最终特征映射为 400 维，进行全连接层的处理。在串联 LSTM 中，句子只进行串联的序列建模。目前，对于问句完整性的识别还没有一个专门的任务，本文人工客服对话数据中的语句补充关系的识别是一个难点，两个实际上不具有补充关系的语句在人工单独看来有可能存在补充关系，这种特点导致了特征的提取与学习有一定的难度。如样例表 5-6 所示，其中在每个对应的序号下，分别给了一个正例和一个负例，‘0’表示两句子之间不具备补充关系，‘1’表示两句子之间具有补充关系。样例 1 是比较容易区分的，但是样例 2、3 中，单从两个句子的角度来看，并不是很容易区分，这是比较困难的。

表 5-6 实验二补充关系识别样例

序号	内容		补充关系
1：	亲我们贴膜 9 号到货就上架的亲	看看有什么问题才可以帮您换的哦	0
	亲我们贴膜 9 号到货就上架的亲	到时候您可以看看	1
2：	又来买肉松	不包邮吗	0
	又来买肉松	40 斤	1
3：	那边快递说找找呢	大概什么时候才会有啊	0
	那边快递说找找呢	主要是现在没有到达记录	1

从表 5-7 的实验结果看，基于 CNN 的并行建模效果最好，其次浅层机器学习的模型结果与串联 LSTM 的结果相差不多。一方面，说明了并行建模的效果要优于顺序建模的效果；另一方面，说明了在不做大量人工特征的情况下，深度模型的效果还是相对优越的。不过深度模型需要大量的数据做支撑，无论是训练的数据上还是词向量的表示上。从两类深度模型的实验看，更大规模的通用字向量，效果更好。在以后的工作中，更深层次的特征还有待进一步的识别。由于本实验中数据存在许多主题非常相近或者在浅层特征上无法简单区分主题的语句，导致在该对话中补充关系的识别难度较大。

表 5-7 基于不同模型的对比实验

实验方法	P	R	F1
SVM	0.578	0.714	0.639
CNN(ty_w2v_char)	**0.633**	**0.731**	**0.678**
Series LSTM(ty_w2v_char)	0.500	0.867	0.634

5.5 补充关系识别在问答匹配中的应用

本节将本章基于并行 CNN 的补充关系识别模型应用到上一章的问答匹配测试中。在上一章节的问答匹配测试中，测试数据中的语句补充关系是由人工标注而来，在实际使用中，我们并不会预先知道问句之间的补充关系。这里，我们将本文中的语句补充关系识别方法和问答匹配关系识别方法相结合，利用补充关系识别方法先对测试数据中问句的补充关系进行识别，然后再利用匹配关系识别方法进行问句与下文的问答匹配。

原先的测试数据由一个完整问句及其 5 条下文所构成。同时，经过数据统计，本对话数据中 98%的具有补充关系的问句跨度不会大于 3 条。因此，补充关系识别在问答匹配中的具体应用过程如下：我们首先使用补充关系识别方法，对该 5 条下文的 3 条上文问句进行问句补充关系的识别，将由基于并行 CNN 的模型识别出的具有补充关系的问句拼接在一起，非补充关系的问句，则单独成为一个问句，以代替原来的完整问句，并依次将新获得的新问句与其对应的 5 条下文进行组合；然后，使用第四章策略一中基于 GRU 的双层 RNN 问答匹配模型进行问句与答句之间的匹配关系识别实验。

表 5-8 补充关系识别与匹配关系识别结合的对比实验

实验方法	MRR(%)	MAP(%)
经问句补充关系识别	**81.94**	**79.42**
未经问句补充关系识别	81.15	78.76

实验结果显示，利用具有完整问句的训练数据训练出的问答匹配模型在未经问句补全的测试数据上进行问答匹配，MRR 达到了 81.15%，MAP 达到了 78.76%。而利用本章基于 CNN 的模型进行问句补全后，问答匹配的 MRR 达到了 81.94%，MAP 达到了 79.42%。不难看出，在实际使用中，本节所提出的补充关系识别与匹配关系识别相结合的方法具有一定的实际意义。

5.6 本章小结

本章主要介绍了语句补充关系任务的相关研究。首先对支持向量机、基于并行 CNN 的模型和串联 LSTM 的模型的实现细节进行了详细的介绍，接着对实验中使用到的三类实验数据的构建作了说明，最后从垃圾数据的过滤、负例构造和词向量的使用等方面对不同方法在不同数据上的结果进行了分析，对话数据中语句补充关系的特征识别具有一定的困难。

结　论

本课题针对交互式问答中的问答匹配关系识别和语句补充关系识别进行了研究。对于问答匹配，进行了社区问答数据和电商客服对话数据的对比实验，构建了基于并行卷积神经网络的模型和基于层次循环神经网络的模型，着重研究深度学习方法在对话数据问答对生成上的效果，并进行对比实验分析。对于语句补充关系识别，采用了基于支持向量机的方法、基于卷积神经网络的方法和基于串联 LSTM 的方法进行对比实验。验证了生成模型在对话数据问答对的提取上的有效性，但对于补充关系识别尚存在一定的困难，最终将补充关系识别和问答匹配关系识别进行了结合。

本课题主要的研究内容包括以下几个方面：

（1）针对中文交互式对话数据中问答匹配关系的识别，提出了基于每轮会话的提取方法和以客服答句为基础基于固定滑动窗口大小的数据提取方法。采用多种方法在对话数据上进行了对比实验，实现了基于 CNN 和基于双层 RNN 的方法，以及在基于 GRU 的双层 RNN 模型上添加注意力机制，改变 RNN 的循环方向。对包括不同的网络结构、不同的数据抽取方式、问句的完整性、问句和答句的信息贡献度进行了实验对比和性能分析。同时对基于 RNN 的不同模型上的阈值选择进行了实验对比分析。验证了在基于每轮对话抽取的数据上，本文实现的双层 RNN 生成模型对于场景信息归纳的有效性，MAP 达到了 84.41%。

（2）针对 Semeval-2016 英文社区问答数据上的问答匹配，使用包含 13967 个词的维基百科英文词向量，采用基于 CNN 和基于 RNN 的方法进行对比，同时对比了相同方法在社区问答数据和对话数据上的性能。由于社区问答数据中存在较弱的上下文场景信息，实验表明，本文中 CNN 匹配的效果要优于 RNN 生成模型的效果。

（3）针对单轮对话数据中，进行问句或答句补充关系的识别。从人工客服对话数据中，抽取具有补充关系的语句以及不具补充关系的负例。实现了基于支持向量机、CNN 模型和串联 LSTM 模型的补充关系识别，分别从数据过滤、负例构造和不同字词向量的角度进行对比实验。单轮对话数据中，补充关系的识别是一个难点。实验结果显示，基于 CNN 的方法表现最优，获得了 0.678 的 F 值。最后，将问句补充关系识别应用于问答匹配。

同时，本文实验仍存在一些需要改进的地方：（1）人工客服对话数据的标

注规模有待提高；（2）在问答匹配任务上，可以考虑加入其它信息的特征；（3）对于语句补充关系的识别，不同轮对话中具有相同主题的语句的非补充关系的判断特征有待进一步识别。对于这些问题，本人将在今后科研的道路上继续探索。

参考文献

[1] Bouziane A, Bouchiha D, Doumi N, et al. Question Answering Systems: Survey and Trends[J]. Procedia Computer Science, 2015, 73(73):366-375.

[2] 郑实福, 刘挺, 秦兵,等. 自动问答综述[J]. 中文信息学报, 2002, 16(6):46-52.

[3] Woods W A, Kaplan R M, Nash-Webber B. The Lunar Sciences Natural Language Information System[J]. Journal of Neuroimmunology, 1972, 174(s 1–2):32-38.

[4] 王宝勋, 刘秉权, 孙承杰,等. 网络问答资源挖掘综述[J]. 智能计算机与应用, 2012, 2(6):54-58.

[5] 吴友政, 赵军, 段湘煜,等. 问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3):1-13.

[6] Bhaskar P, Pakray P, Banerjee S, et al. Question Answering System for QA4MRE@CLEF 2012[C]//The Workshop on Question Answering for Machine Reading Evaluation. 2012:1-11.

[7] Kando N. Overview of the Fifth NTCIR Workshop[C]//Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. 2005:2-90.

[8] 董燕举, 蔡东风, 白宇. 面向事实性问题的答案选择技术研究综述[J]. 中文信息学报, 2009, 23(1):86-94.

[9] Kelly D, Lin J. Overview of the TREC 2006 ciQA Task[J]. ACM SIGIR Forum, 2007, 41(1):107-116.

[10] Zheng Z. AnswerBus Question Answering System[C]//International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. 2002:399-404.

[11] Kupiec J. MURAX: A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001:181-190.

[12] Hovy E H, Gerber L, Hermjakob U, et al. Question Answering in Webclopedia[C]// The Ninth Text REtrieval Conference. 2000, 52: 53-56.

[13] Feng D, Shaw E, Kim J, et al. An Intelligent Discussion-bot for Answering Student Queries in Threaded Discussions[C]//Proceedings of the 11th

International Conference on Intelligent User Interfaces. ACM, 2006: 171-177.

[14] Huang J, Zhou M, Yang D. Extracting Chatbot Knowledge from Online Discussion Forums[C]//Proceedings of the International Joint Conference on Artificial Intelligence. DBLP, 2007:423-428.

[15] Cong G, Wang L, Lin C Y, et al. Finding Question-answer Pairs from Online Forums[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 467-474.

[16] Ding S, Cong G, Lin C Y, et al. Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums[C]//Proceedings of the Meeting of the Association for Computational Linguistics. 2008:710-718.

[17] Cao Y, Yang W Y, Lin C Y, et al. A Structural Support Vector Method for Extracting Contexts and Answers of Questions from Online Forums[J]. Information Processing & Management, 2009, 47(6):514-523.

[18] Wang B, Liu B, Sun C, et al. Extracting Chinese Question-answer Pairs from Online Forums[C]//IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2009:1159-1164.

[19] Heilman M, Smith N A. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1011-1019.

[20] Severyn A, Moschitti A. Automatic Feature Engineering for Answer Selection and Extraction[C]//Conference on Empirical Methods in Natural Language Processing. 2013, 13: 458-467.

[21] Yih W T, Chang M W, Meek C, et al. Question Answering Using Enhanced Lexical Semantic Models[C]//Meeting of the Association for Computational Linguistics. 2013:1744-1753.

[22] Wang D, Nyberg E. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering[C]//Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. 2015:707-712.

[23] Hu B, Lu Z, Li H, et al. Convolutional Neural Network Architectures for Matching Natural Language Sentences[J]. Advances in Neural Information Processing Systems, 2015, 3:2042-2050.

[24] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv:1409.0473, 2014.

[25] Sainath T N, Vinyals O, Senior A, et al. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015:4580-4584.

[26] Feng M, Xiang B, Glass M R, et al. Applying Deep Learning to Answer Selection: A Study and an Open Task[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015: 813-820.

[27] dos Santos C, Barbosa L, Bogdanova D, et al. Learning Hybrid Representations to Retrieve Semantically Equivalent Questions[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015, 2: 694-699.

[28] Shang L, Lu Z, Li H. Neural Responding Machine for Short-Text Conversation[C]// Neural Responding Machine for Short-Text Conversation. Association for Computational Linguistics, 2015:1577-1586.

[29] Zhou X, Hu B, Chen Q, et al. An Auto-Encoder for Learning Conversation Representation Using LSTM[C]//International Conference on Neural Information Processing. Springer International Publishing, 2015: 310-317.

[30] Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about entailment with neural attention[C]//International Conference on Learning Representations. 2015:202-211.

[31] Tan M, Santos C D, Xiang B, et al. Improved Representation Learning for Question Answer Matching[C]//Meeting of the Association for Computational Linguistics. 2016:464-473.

[32] Li J, Monroe W, Ritter A, et al. Deep Reinforcement Learning for Dialogue Generation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016:1192-1202.

[33] Wen T H, Vandyke D, Mrksic N, et al. A Network-based End-to-end Trainable Task-oriented Dialogue System[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2016:438-449.

[34] Cortes C, Vapnik V. Support-vector Networks[J]. Machine Learning, 1995, 20(3):273-297.

[35] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[36] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word

Representations in Vector Space[C]//Proceedings of Workshop at International Conference on International Conference on Learning Representations, 2013:386-398.

[37] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]//International Conference on Neural Information Processing Systems. 2012:1097-1105.

[38] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Proceedings of Empirical Methods in Natural Language Processing, 2014: 1746-1751.

[39] Lipton Z C, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning[J]. arXiv preprint arXiv:1506.00019, 2015.

[40] Bengio Y, Simard P, Frasconi P. Learning Long-term Dependencies with Gradient Descent is Difficult[J]. IEEE Transactions on Neural Networks, 2002, 5(2):157-166.

[41] Hochreiter S, Schmidhuber J. LSTM can Solve Hard Long Time Lag Problems[J]. Advances in Neural Information Processing Systems, 1996:473-479.

[42] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2014:1724-1734.

[43] Chung J, Gulcehre C, Cho K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[44] Mnih V, Heess N, Graves A. Recurrent Models of Visual Attention[C]//Advances in Neural Information Processing Systems. 2014: 2204-2212.

[45] Kusner M J, Sun Y, Kolkin N I, et al. From Word Embeddings to Document Distances[C]//In Proceedings of the 23th International Conference on Machine Learning, 2015: 957-966.

[46] 户保田. 基于深度神经网络的文本表示及其应用[D]. 哈尔滨工业大学, 2016: 41-47.

[47] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[48] Sutskever I. Training Recurrent Neural Networks[D]. University of Toronto,

2013:74-77.

[49] Nakov P, Màrquez L, Moschitti A, et al. SemEval-2016 Task 3: Community Question Answering[C]//International Workshop on Semantic Evaluation. 2016:525-545.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《交互式问答中的语句关系识别方法》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：陈静　　日期：2017 年 7月 3 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：陈静　　日期：2017 年 7 月 3 日

导师签名：　　日期：2017 年 7 月 3 日

致　谢

时光荏苒、岁月如梭，两年的硕士研究生活即将结束了，从当初刚入学的懵懂新人到如今科研道路上的探索者，除了不舍和期待之外，更多是内心对于研究生道路上曾经帮助过我的师长同学朋友的感恩之情。

感谢我的导师陈清财教授。陈老师严谨的科研作风和乐观的生活态度深深影响了我。无论是算法问题还是大方向的技术问题，陈老师总能及时的给予我深刻的指导，深入浅出的为我剖析、解惑；在一起讨论学术问题的过程中，他渊博的专业知识和活跃的思维，总能迸发出不同的创新点，在我的课题研究上给了我很大的帮助，同时也培养了我的创新意识，为以后的工作打下一定的基础。此外，陈老师随和的为人，谦虚的学习态度也深深影响了我，为我树立了一个好的榜样，令我受益匪浅。同样感谢医疗组的汤步洲副教授，每当在课题的研究上遇到困难时，汤老师总能给出宝贵的指导意见，让我在技术上不断有了新的认识和提升。同时汤老师良好的科研工作习惯也深深感染了我，培养了我认真的科研态度和良好的学术研究习惯，少走了很多弯路。非常感谢两位老师在学术上给我的指导和鼓励。

感谢博士师兄刘欣，师兄在课题研究的技术上，给了我许多专业的指导，师兄带领我走进了深度学习的大门，教我分析问题，分析模型。无论是在技术上还是科研方法上，都教会了我许多。同时也感谢博士师兄周小强、博士师姐吴湘平、博士师兄陈凯、硕士师兄庄烈彬，在我完成毕业课题的过程中，给了我许多鼓励和帮助，帮我融入实验室这个大家庭，让我不断勇往直前，同时也感谢其他师兄师姐对我的各种帮助和照顾，促进了我的进步。

感谢 15 级的小伙伴们，胡江鹭、吴宇航、郑志辉、熊思兰、张纪绪等同学，在两年的研究生经历中，无论在学习上还是生活上，他们都给了我很多的关心和支持，让我感受到了同学情的温暖。同时，也感谢我的好朋友们，在我情绪低落的时候给我安慰，给了我精神上很大的支撑。

感谢一路陪伴我的亲人，他们永远是我的港湾和坚强的后盾，没有他们的养育、包容、理解和支持，也就没有今天的我，感谢他们一直让我做自己想做的事情。

最后，要感谢给我提供广阔学习平台的母校，给了我学习的机会和良好的环境，感谢智能计算研究中心给了我优越的科研条件，为我的论文顺利完成奠定了基础。

阅读此文的还阅读了：

1. 语句运用的策略与方法
2. 扩展语句解题方法指津
3. 以规则为主的英语句子边界识别方法的C#实现
4. 一种语句识别方法与系统
5. 交互式问答专利技术综述
6. 扩展语句的四种方法
7. 句子、语句与言语行为关系探析
8. 解读循环语句?识别两类模型
9. 句子、语句与言语行为关系探析
10. 问答社区中的问答处理方法及问答系统
11. 论俄语问答式对话统一体反应语句的繁化
12. 藏语句子边界识别方法
13. 网络考试中的人脸识别方法研究
14. 论现代汉语中的连谓语句和状中语句的区分方法
15. 论电子商务中的交互式关系营销
16. 亲子关系问答
17. 汉语兼语句及其在英语、维语中的对应关系
18. 机械手在交互式手语识别系统中的解析应用
19. 脑卒中的早期识别方法
20. 交互式问答的关系结构体系及标注
21. 句子、语句与言语行为关系探析
22. 一种邻区关系识别方法
23. 交互式问答中的语句关系识别方法
24. 基于用小波变换改进MFCC的语句识别方法
25. 搭配识别中的统计方法

26. 框架关系推理在阅读问答中的应用

27. 交互式问答中的指代消解研究

28. 汉语语句相似度算法在问答系统中的应用研究

29. 物理解题中的识别方法

30. 语句识别装置、语句识别方法、程序和媒体

31. 防空识别区知识问答

32. 汉语句法中的框-梹关系

33. 汉语句子谓词的自动识别方法研究

34. 一种注采连通关系识别方法的研究

35. 英语句法关系分析

36. 用于交互式演示系统中的手指识别方法

37. “望、闻、问、切”方法在植物分类识别中的运用

38. 语句扩展的方法

39. 自动识别语句关系和实体的方法及装置

40. “扩展语句”的两种方法

41. 交互式问答系统中待消解项的识别方法研究

42. 一种注采连通关系识别方法的研究

43. 面向技术用途的关联关系识别方法研究

44. 车牌识别中的图像定位及分割方法

45. 面向交互式问答的人物事件关系抽取方法研究

46. 汉语句法中的框-梹关系

47. 语句扩展方法漫谈

48. 交互式问答系统中的省略恢复研究

49. 俄语句际关系中的实际切分与交际接应

50. 基于排序方法的汉语句际关系树自动分析