硕士学位论文

网络购物环境下的问句答案匹配方法研究

RESEARCH ON QUESTION AND ANSWER

MATCHING ORIENTED ONLINE SHOPPING

陈俊杰

哈尔滨工业大学

2013 年 06 月

国内图书分类号：TP 391.3　　学校代码：10213
国际图书分类号：621.3　　密级：公开

工学硕士学位论文

网络购物环境下的问句答案匹配方法研究

硕士研究生：　陈俊杰
导　　师：　王晓龙 教授
申请学位：　工学硕士
学科、专业：　计算机科学与技术
所在单位：　深圳研究生院
答辩日期：　2013 年 06 月
授予学位单位：　哈尔滨工业大学

-

Classified Index: TP 391.3
U.D.C: 621.3

Dissertation for the Master Degree in Engineering

RESEARCH ON QUESTION AND ANSWER MATCHING ORIENTED ONLINE SHOPPING

Candidate: Junjie Chen

Supervisor: Prof. Xiaolong Wang

Academic Degree Applied for: Master Degree in Engineering

Specialty: Computer Science and Technology

Affiliation: Shenzhen Graduate School

Date of Defence: Jun, 2013

Degree-Conferring-Institution: Harbin Institute of Technology

摘　要

网络购物已经成为人们生活中不缺少的购物方式。它具有方便、快捷等特点，使用户能够足不出户浏览和购买想要的商品。人们通过网络会话的方式向客服咨询商品信息。客服通常会同时回答多个用户提问的问题，导致服务质量差，容易使用户流失。如果有一个辅助问答系统帮助客服检索信息，并给出建议答案，将大大提高客服的服务效率与质量。问答系统在网络购物咨询方面具有广阔的应用前景。

问答系统的效果通常依赖于知识库的规模及质量，因此从网络购物记录中提取问答对是构建整个系统的核心问题。在网络购物记录中存在多问句与多答案交叉的复杂对应关系，其最大特点就是答案的滞后性。用户连续提出多个问题，客服逐一的回答，问题和答案可能不是相邻且一一对应的。目前，知识库中的问答对多都是人工从复杂对应的关系中提取问答对，不仅费时费力、维护成本高，而且不能实时更新。

为解决这一问题，本文把问句答案匹配的判断作为一个二分类任务，根据语料特点，设计了三个分类方法：基于特征匹配的方法是利用问句与答案中的句式类型、公共词序列、概念关系三个特征判断是否为匹配的问答对；基于冗余信息的方法是利用现有问答系统的检索功能，计算检索答案与候选答案相关度判断是否为匹配的问答对；基于词共现的方法是统计问句与答案中共同出现的词汇对，计算词汇的相关度来判断问句与答案是否匹配。

对三个分类方法分别设置相应实验，结果表明它们都能有效的从网络购物记录中提取问答对。最后，本文将三个分类方法有机的组合起来，形成一个自训练模型框架。该框架能够利用少量的标注语料及大量未标注的语料迭代训练，从中提取问答对。经过多次迭代训练，自训练模型的准确度明显高于三个单一的分类方法。

关键词：网络购物；问答匹配；问答系统；自训练模型

Abstract

Online shopping has become a necessary in people's lives. It is convenient and fast, allowing users to browse and shopping in home. However, online shopping customer service often chat with multiple user and reply a lot questions at the same time, resulting in the loss of customers. If there is an assisting question answering system help customer retrieve information and gives some suggestions for the answer, it will improve the efficiency and quality of service customer service greatly. Question answering system has wide application in online shopping consultation.

The performance of a question answering system depends on the knowledge base's quantity and scale. So, it has become a core problem extracting from chat log. There are a number of questions corresponding with a number of answers in the corpus of network session. The most obviously feature of the complex corresponding is the answer lag. Buyers raised a number of questions, customers service answer them one by one. Questions and answers may not be adjacent. At present, most of the question answering system knowledge base of question-answer pairs are artificial constructed, which not only is time-consuming and high maintenance cost, but also cannot update in real time.

To solve this problem, we propose a self-training model framework that extract question answering pairs from a large chat records. The framework can use a small labeled corpus and a mount unlabeled corpus to iterate training model. In this framework includes three different models: the feature detection based matching evidences model which include sentence type, common word sequence, the concept of the relationship; the redundant information based relation model which compute similarity between the searching relevance answers and the candidate answer; the word co-occurrence based relation model which statist the common vocabulary sequence between question and answer.

In the corresponding experiments, test each model and analysis their results. Three models are effective for extracting question-answer pairs from session

record. Finally combine three models together, forming a self-training framework. The accuracy of the framework is greatly improved after the combination.

Keywords: online shopping, question and answer matching, self-training framework

目录

第1章 绪 论

1.1 课题背景目的和意义

随着信息大爆炸的出现，基于互联网搜索引擎的检索方式不能满足人们对快速、准确地获取信息的需求。因此，问答系统作为高级的信息检索形式，逐渐受到用户的喜爱。它是一个自动机[1]，能够回答用自然语言提出的问题，并且输出一个简洁、准确的答案或者候选答案列表，而不是一堆的相关文档。用户只需要用口语化的方式直接提问即可，不必思考该使用什么样的问法才能够得到理想的答案，并且不需要逐个查看搜索引擎返回的网页，这样就提高信息的查找效率。

问答系统使用了大量的自然语言处理技术，如词性标注、句法分析、信息检索和答案抽取等，其中有些系统[2]使用了复杂的逻辑推理[3]技术。自从1993 年，美国 MIT 人工智能实验室开发出全世界第一个基于 Internet 的问答系统——START，到目前为止，问答系统已经取得了很大的发展，世界主要语言都有问答系统，甚至有的支持多种自然语言。现在国际上，问答系统的是一个热门的研究方向，很多研究机构都积极参与到该领域的研究。

全国的网络购物注册人数早已突破数亿大关。网上购物逐渐改变了人们的购物习惯，尤其对年轻一代，越来越成为生活中不可缺少的一部分。网络咨询的方式使人们能够更详细、方便的了解网上购物的商品信息。当人们在享受生活便捷的同时，可能网站的每一个客服都正在忙碌着应对多个用户提出的问题。一个商家每天都要接受大量的用户咨询，人们通常就一个或多个商品提问，由客服为其回答，并且用户一般都针对常见的问题咨询，这些问题的重复率很高。客服的疲惫导致服务质量差，容易使用户流失。如果有一个辅助问答系统帮助客服检索信息，并给出建议答案，将大大提高客服的服务效率与质量。辅助问答系统通常包括三个部分：问题分析，信息检索和答案提取。其中，信息检索的效果需要依赖知识库的规模和质量，因此知识库的建立对问答系统的结果有着至关重要的影响。

基于网络环境的会话语料中，问句与答案并不是一一对应的，存在多问句与多答案交叉的复杂对应关系。复杂对应的最大特点就是答案的滞后性。因此必须从复杂的会话记录中提取有效的问答对构建知识库。现有的辅助问答系统知识库通常都是通过人工根据已有的历史会话记录构建。人工从会话

记录中提取问答对的方法费时费力，并且不能及时的处理新产生的会话记录。因此，设计一个自动从会话记录中提取出有效问答对的计算方法是一个亟待解决的问题。

1.2 国内外研究现状

1.2.1 问答系统研究现状

自从 1960 年代人工智能研究的起步阶段，人们就提出设计一个问答系统，让计算机能够回答用自然语言提出的问题。1961 年成功的设计出了一个问答系统用于回答单季美国职棒大联盟相关比赛问题[4]。随后出现了大量的问答系统，其中多数是面向专业领域的专家系统，涉及到心理学、医学和自然科学等等。

网络和搜索技术日新月异，人们对信息的需求也变得更加迫切。搜索引擎的发展大大满足了人们对信息获取的需求，比较有名的搜索引擎有 Google、Baidu、Yahoo 等。用户只需输入一些关键字，搜索引擎都能快速的返回相关的网页。目前传统的搜索引擎存在很多不足的地方，主要体现在两个方面：

（1）返回相关网页，而不是直接的答案。传统的搜索引擎根据用户输入的关键字返回一系列按相关度排序的网页，用户需要逐个查看搜索引擎返回的结果，大大降低了信息查找的效率。

（2）基于关键词的逻辑组合不足以表达检索需求。复杂用户的检索需求很难用逻辑组合来表达，用户无法清楚的表述自己的检索意图，致使搜索引擎返回的相关网页也无法满足用户需求。

互联网的普及，互联网上的信息越来越丰富及新的自然语言处理技术和方法的提出，促使问答系统研究进入一个新的阶段。START 是全世界最早的基于 Internet 的问答系统之一（http://start.csail.mit.edu/）。由美国麻省理工学院人工智能实验室于 1993 年开发出，该系统一直运行，目前能够回答地理、科技、文化、历史都基本问题。多语种的自动问答系统 AnswerBus[5]不仅可以回答英语的问题，还可以回答法语、西班牙语等六种语言提出的问题（http://www.answerbus.com）。Kupiec 等[6]设计了基于百科全书人常识问答系统 MURAX，它可以回答一般性的知识问题。FAQFinder[7]则以“问题-答案对”为基础，通过基于向量空间的搜索引擎从已知的问题-答对中获取答案。

越来越多的问答系统面世，但由于各个系统应用于不同领域，加上系统复杂度高，很难做客观的评估与比较。问答系统的评估是非常耗时耗力的，首先要生成一个包含大量问题的测试集，同时以人工方式把每一个问题可能对应的答案从比赛语料中挑选出来。问答系统比赛的出现促进问答系统的健康发展，并提供一个统一比较、评测的平台。TREC（Text Retrieval Conference）是国际上著名的文本检索会议，早在 1999 年就设立 QA Track，其目标是对面向开放域基于大规模文本库的自动问答系统评测[8-12]。2003 年由日本国立情报学研究所 NII 第一次举办日文的问答系统比赛——NTCIR 会议（NTCIR Workshop）[13-16]；同样，欧洲于 2003 年由 CLEF（CrossLanguage Evaluation Forum）主办欧洲语言的问答系统比赛[17-20]，内容最为丰富，其中参与比赛的就达九种语言，另外还有跨语言的问答比赛。虽然中文是世界上第二大语言，但直到 2005 年才由 NTCIR 第一次举办有关中文的问答系统比赛。中文问答系统的准确率仍然与英文及其他语言有较大的差距[21]。目前最引人瞩目的英文问答系统 Watson 是由 IBM 研究团队历经四年紧张的研发，于 2011 年 2 月在《Jeopardy!》智力竞赛中亮相，凭借其优秀的自然语言处理技术和极快的运行速度，与《Jeopardy!》竞赛节目中最为著名和最为成功的两位冠军选手 Ken Jennings 和 Brad Rutter 一决胜负，最终取得冠军[22-24]。

问答系统通常包括三个主要部分：问题分析、信息检索和答案提取。如图 1-1 所示：

当用户以自然语言的方式输入自己的问题，问题分析是整个问答系统处理的第一步，首先要对其进行解析，将用户输入的问题转换成问答系统易于检索的表示形式。其准确率直接影响着后面的模块的结果。

信息检索模块作用是从已有的文档库或开放网络中检索相关文档，找出相关度比较高的信息。其准确度主要取决于文档库的规模及质量。

答案提取模块从返回的一堆相关文档中提取出候选答案列表。最后通过答案确认找到一个简洁、准确的答案。答案抽取模块的效果直接影响用户的体验。

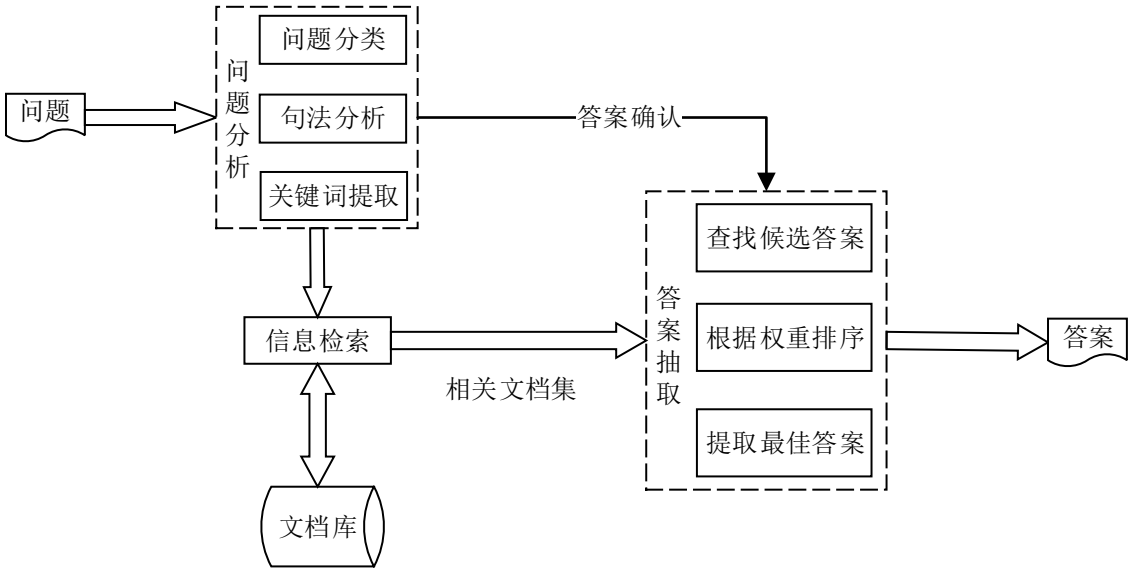

图 1-1 问答系统结构

1.2.2 问答系统分类

问答系统技术涵盖了文本处理技术、自然语言处理技术、信息提取技术、搜索引擎技术等多个方面的技术，可以从知识领域、答案来源等角度来分类。

（1）基于问答系统的知识领域分类。依据问答系统知识领域的开放性不同，可分为“封闭领域”以及“开放领域”两大类系统。

封闭领域系统的知识领域是一个限定的、与外界无交换的系统。最典型的封闭系统是专家系统，其专注于回答特定领域的问题，如医药、心理咨询或特定活动等。由于知识领域受限，封闭领域系统有比较大的优化空间，可以加入领域专属本体知识库[25]、句式模版、推理规则等，或将答案来源全部转换成结构化数据来有效提升系统的准确率。这些系统的共有缺点是：有着很大的局限性。关于专业领域内的问题回答的准确率很高，一旦涉及到开放性的领域问题，系统的准确度就会大幅下降。系统规模较小、缺少大规模真实数据以及问题有局限性、需要大量的人工维护数据等等一系列现象都是专家系统难以克服的问题。

开放领域问答系统不对问题内容设定特殊范围，知识领域也多种多样涉及到各个方面，并且数据可能实时更新。目前开放领域的答案来源多是互联网等开放性的大规模文本数据。开放领域问答系统的准确度远不如封闭领域的问答系统。在开放领域中，由于数据来源格式不一样，并且不是结构化的形式，并且不同来源提供的数据权威性不同，并没完成的经过人工验证，所

以开放领域问答系统的准确度较低。目前大多数系统回答的问题类型主要是基于事实的简单句，比如：时间、地点、人物、历史事件等。对于非事实类的问题，回答效果总是不尽人意。最早的基于网络开放问答系统 START，从 1993 年开放至今仍提供服务。现在该系统只能回答英语提问的基本的简单的问题。

（2）基于问答系统的答案来源分类。依据问答系统处理的数据格式不同，可将其分成：基于结构化数据的问答系统和基于自由文本数据的问答系统[26,27]。

第一种类型是基于结构化数据的问答系统。基于结构化数据的问答系统的主要思想是通过分析问题，把问题转化为一个类似数据库的查询，然后在结构化数据库中进行搜索，返回的查询结果即为问题的答案。由于在查询的过程中，答案的查找是基于数据的高度匹配得到的，所以对于答案提取的过程并不重要。构建一个基于结构化数据的问答系统有两个关键问题：一是需要构建一个比较完备的结构化数据库；二是高效、准确地把问题转化为数据库形式的查询。系统结构如图 1-2 所示：

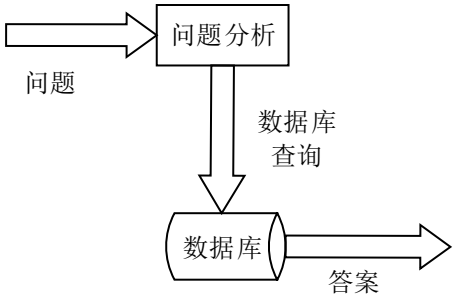

图 1-2 基于结构化数据问答系统体系结构图

基于结构化数据的问答系统大量出现在上世纪人工智能和计算语言学兴起的时期。目前，聊天机器人、基于知识库的问答系统都是基于结构化数据的问答系统。聊天机器从是模拟人的用语习惯，与人进行人性化的自然语言交流。ALICE 是由 Wallace 开发的一个开源的聊天机器人，连续获得 2000-2002 年“Loebner Prize”比赛冠军[28]。它使用 AIML 语言表示知识，并定义了丰富的标签及大量模板，但是并没复杂的算法。ALICE 主要的答案来源于结构化的聊天记录，AIML 语言提供内联机制，可以方便地将多个知识库合并起来。基于知识库的问答系统一般都是联合查询多个知识库，并利用推理技术，来分析和理解用户问题。一般说来，知识库的数量与质量、答案推理的

准确性，是其性能的决定性因素。大多知识库都是基于本体构建的，当前本体库的建立采用手工方式，并且开发原则、设计标准都不一样，建立知识库远远未成为达到工程性活动。主要基于知识库的问答系统有专家系统、常用问答系统（FAQ）。

第二种类型是基于自由文本的问答系统。自由文本是指用自然语言组成的文本，表达方式灵活多样，与结构化数据不同。即使对同一类事件，可以有很多种不同语言表述方式，因此很难直接从未经任何处理的自由文本中提取信息[29]。基于自由文本检索系统的核心是对自由文本的处理，其中包括文本处理技术、信息检索技术、信息抽取技术等自然语言处理技术。信息抽取技术的优劣直接影响到对系统回答性能的评测结果。系统结构如图 1-3 所示：

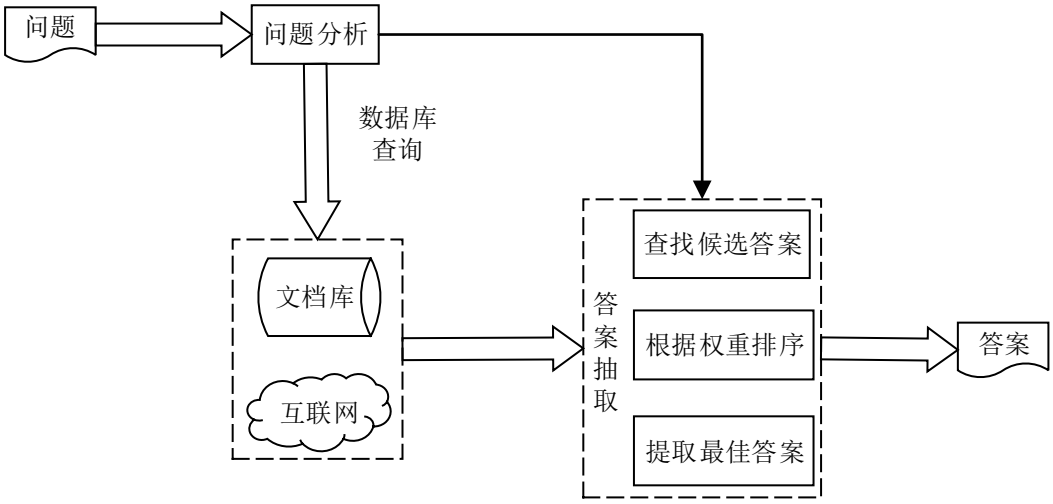

图 1-3 基于自由文本问答系统体系结构图

这类系统一个较大的挑战在于：由于自然语言的表达具有很大的灵活性，问题和候选答案可以用多种不同的方式进行表述，也就是说，问题与候选答案之间的词汇不一定是匹配的，而是存在着语义的关联[30, 31]。为了解决这个问题，最初，一般都采用句模匹配的方法。该方法在问题类型确定的前提下，依据答案中含有的上下文信息，激活与此类型问题相关的答案模板用来提取答案。此种方法一般都是在小规模的语料中提取出精确的信息，切换到一个新的领域中，需要用户创建大量新的规则或手工标注新的训练语料。近年来，随着机器学习的方法改进，逐渐成为主流。机器学习的答案抽取方法一般是基于这样一种假设：含有正确答案的句子与问句的距离应该小于未含有正确

答案的句子与问句的距离。具体方法是，把问句和候选答案映射到一个设定的空间，计算问句和候选答案在此空间的距离。此外还可以利用语义词典，引进词汇级别的推理技术以及逻辑层面的推理技术。目前提取出信息准确性仍然不高。

通过用自由文本集数据而非传统的知识库，此类系统能够大大减少搭建和维护更新知识库所消耗的人力和物力，并解决了知识库规模无法满足的难题。运用信息抽取技术，只将确定的答案信息提交给用户，而不是返回相关的文档或者网页，避免用户在获得的大量网页内容或文档信息中二次搜索需求答案而浪费的时间。

1.2.3 问答对提取技术的研究

问答对提取是从在线论坛、电子邮件、即时通讯记录等资源中挖掘问答对，并作为问答系统的语料。目前的大部分的解决方法主要是问答对提取作为一个二分类问题，通过文本特征及各种特征的组合，使用分类或排序算法来预算答案。

Huang 等人利用分类及提出级联框架方法[32]（Cascade Based）从在线论坛中提取问答对，对于从英文在线论坛中提取问答对做了开创性的工作。Feng 等人利用余弦相似度匹配用户查询词及回复信息[33]，实现了多线索的讨论机器人。Gao 提出一个基于序列模式（Sequential Patterns Based）的分类方法、基于图的传播方法从论坛中提取问题和答案[34]，并取得良好的实验结果。Ding 等人所作的研究引入条件随机场模型来抽取答案和上下文[35]，研究证明跳跃链条件随机场（Skip-chain CRFs）和二维条件随机场（2D CRFs）的结合可以大约 70%的准确度从论坛中提取给定问题的上下文和答案，但同时也以牺牲时间为代价。其后，Yang 等人在 Ding 等人的基础上，发掘了更多可能存在的句子间相互关系，采用结构化的支持向量机（M-SVM）方法来抽取给定问题的答案和上下文[36]，相对于决策树方法和 B-SVM 的方法，有了明显的改进，该项研究拓展了潜在问答规律，但是准确率仍然没有得到大步提高。目前，哈尔滨工业大学 Wang 等人对如何从中文在线论坛中识别问答对进行了研究[37]，他们提出一个基于序列规则的方法识别问题和基于论坛结构的非文本特征方法改善识别效果。

1.3 本文主要研究内容与组织

1.3.1 本文内容

在基于知识库的辅助问答系统中，知识对的收集是重要的环节。因此知识库的规模和质量是影响问答系统性能的重要因素。目前所使用的知识库是由人工构建的，在规模和效率上都受到很大的制约。

本文提出一个提取网络购物记录中问答对的自训练模型框架。通过分析网络购物语料中的语言现象，设计了三种计算方法：基于特征匹配的相关度计算方法、基于冗余信息的相关度计算方法、基于词共现的相关度计算方法。最后将这三种计算方法组合起来，构建成一个提取会话记录中问答对的自训练模型框架。具体研究内容如下：

（1）详细统计分析网络购物语料中语言现象。本文收集了大量的网上购物的会话语料，首次对其中的语言现象进行了详细统计，重点对问句答案匹配现象进行了分析和讨论。

（2）设计基于特征匹配的相关度计算方法。分析网络购物语料中问句与答案对应的特征，根据相应特征提取问句与答案匹配证据，并以证据个数的多少来判断给定的问句与答案是否匹配。

（3）设计基于冗余信息的相关度计算方法。依据问答系统的知识库中存在的大量冗余信息，并利用现有的问答系统检索机制，在系统中搜索给定的问题，得到一系列相关答案。计算这些相关答案与候选答案的相关度，来判断给定问题与候选答案的匹配度。

（4）设计基于词共现的相关度计算方法。把一个问答对作为一个共现窗口，对标注语料进行词共现频率统计，得到词共现计算方法，计算给定问题与候选答案中有意义词的相关度，来判断两者的匹配程度。

（5）构建提取会话记录中问答对的自训练模型框架。分析三个计算方法的特点，将它们有机的组织起来，相互监督，构建成一个提取会话记录中问答对的自训练模型框架。

1.3.2 本文的组织

第 1 章为绪论部分，主要阐述了本课题的背景、目的和意义，重点部分是介绍问答系统、问答系统的分类以及问答对提取在国内外的发展情况。最后简明的概述了本文的主要研究内容和组织安排。

第 2 章是对问句答案匹配的问题描述，先描述了本课题所要解决问题，

并介绍了问句答案匹配的语言现象，然后对要解决的问题进行形式化描述。

第 3 章是问句答案匹配方法的研究。首先是基于特征匹配的计算方法，对特征选择作详细描述，特征主要包括：句式类型、公共词序列、概念关系对。描述了特征检测的算法及基于其分类的过程。然后是基于冗余信息的相关度计算方法，简述算法的流程：问题复述、问答系统检索、句子相关度计算、问句答案匹配度计算。接着阐述了基于词共现的相关度计算方法，重点论述训练策略，提出“慢增快减”策略。

第 4 章语料的分析，先介绍了本课题对收集到的语料的一些基本上信息，概括说明了语料中的非规范语言现象，重点对问句答案匹配现象进行标注分析。最后说明训练集及测试集的收集，并统计分析了语料中的有效问答对比例。

第 5 章是对前面三个计算方法实验的结果及对比分析。首先对比分析了三个计算方法的特点，然后基于这三个计算方法构建自训练框架，将它们三个组合起来进行实验，并给出结果分析。

第2章 问句答案匹配的问题描述

2.1 引言

网络购物的记录中存在多问句与多答案交叉的复杂对应关系。在解决这一问题前，有必要对其中的概念做出阐述，及对要解决的问题进行形式化描述，以便在第 3 章中更方便的介绍具体的计算方法。

2.2 问句答案匹配的相关概念

以真实的网上购物聊天语料为基础，对用户和客服的会话记录进行统计分析。语料中有些会话是简单的一问一答规整的对应，这种情况我们称之为简单对应；有些会话是多个问句，多个答案，问句与答案交叉对应，这种情况我们称之为复杂对应。简单对应中的问句答案本身就是匹配的。复杂对应中问句与答案的匹配有以下几种情况：

一对多的匹配关系，即一个问句，多个答案与之匹配；多对一的匹配关系，即多个问句，一个答案与之匹配；多对多的匹配关系，即多个问句，多个答案与之匹配。

本文主要研究网络购物语料中复杂对应关系下的问句答案匹配。语料中复杂对应的最大特点就是答案的滞后性。在真实的对话中，用户连续提出多个问句，客服逐一的回答，所以问句和答案可能不是相邻且一一对应的。

例 2-1 会话记录中一个例子：

Q1：今天能发货吗？

A1：亲 我们是今天发货的

Q2：如果不发申通，还能发什么快递

Q3：你能不能给我包邮啊

A2：我们的默认为申通及韵达快递

A3：可以包邮的，亲

在例 2-1 中，Q2 与 A2 为一组问答对，Q3 与 A3 为一组问答对，但是它们不是相邻，而且交叉出现的。为了表述方便，首先定义如下：

定义 2-1 会话：说话双方从说话开始与结束的整个对话。

定义 2-2 话轮：一个说话者在会话过程中从开始说话起到讲完停下或被对方强行打断为止，所说的一段话。在例 2-1 中，Q1 为一个话轮；Q2，

Q3 合为一个话轮。

定义 2-3 语句：一个话轮中包含一个或多个分句，每一个分句称作一个语句。在例 2-1 中，Q2，Q3 合为一个话轮，其中 Q2，Q3 分别为一个话句。

定义 2-4 交叉匹配：由于网络聊天的回答具有滞后性，说话一方提出问句后，对方并不是立即回答。大多数情况下，回答的答案与问句并不是相邻的，当问句与答案不是相邻语对时称作交叉匹配。在例 2-1 中，Q2 与 A2，Q3 与 A3 是问答对，但问句与答案不是相邻的，而是交叉出现。

在会话分析理论中，会话结构的最小单位“话轮”。Sacks 和 Schegloff 通过研究了大量口语会话素材后提出了具有深远影响的“话论转换”理论，一篇连贯的会话至少包括发话人和听话人双方各自发出一个话轮，会话的参与者互相配合通过“话轮转换规律”交替发话[38]。他们把由不同的说话人各自所说的话轮交替组合称为“话轮对”。其中那些位置紧邻的“话轮对”的上句和下句经常固定地配对出现，语义上往往呈现出一定的联系。“话轮对”又称为相邻对，即在两个人的会话中，一个人发话，另一个有对此做出回应，两者的话语构成一对意义上匹配的语列，其中第一个人的话语为“引发语”，第二个人的话语为“应答语”。

问答对是相邻语对的一种，Sacks 和 Schegloff 指出相邻语对有以下 5 个特征：

a) 由两个话轮构成;
b) 两个话轮相邻接;
c) 两个话轮由不同的说话人说出;
d) 第一部分称为引发语，第二部分称为应答语，第一部分在第二部分之前;
e) 引发语的类型影响到应答语的选择。

在网络购物语料中，一个会话包含多个话轮转换，一个话轮包含多个语句。相邻语对匹配是指话轮中的一个语句与相邻话轮中一个语句匹配，即两个话轮中的语句是交叉匹配的。本文把相邻语对称作常见的广义问答对，下文提到的问答对不作特殊说明都是指相邻语对。

一般情况下，用户在前一个话轮中提出问题，客服在下一个话轮中给予回答。特殊情况下，用户的问题并不是在相邻的话轮里给出回答，而是在以后隔着多个话轮才做出回答。

本文所讨论的问句答案匹配是检测相邻话轮中的问句与答案是否为一对意义上匹配的问答对，对于不相邻话轮中问句答案匹配情况不做处理。

2.3 问句答案匹配现象

本文将收集的网络购物语料中问句答案匹配分为显性匹配和隐性匹配，其中显性匹配主要指通过问句与答案中一些词的特征现象来匹配，隐性匹配主要指语义级别的匹配。

2.3.1 显性匹配

显性匹配是指引发语与应答语中包含明显匹配的词汇和句法特征，根据这些特征就能识别两者的匹配关系。

（1）**指代匹配**　　如果在相邻语对的应答语中包含代词，那么代词常常指代引发语中出现的词语，表示该引发语与应答语论述相关的内容。代词包括指示性代词（这，那），指称性代词（你，他，它），指代词组（那个红色的）。

例 2-2 含有指示性代词的例子：

Q：老板，我要一个透明的手机壳

A：那款卖完啦

（2）**概念关系匹配**　　每个概念都有对应的概念集，通过刻画这些对应关系来描述一个概念的性质。问句与答案中包含匹配的概念对，说明论及相同内容。

例 2-3 概念匹配的问答对：

Q：短袖都有什么颜色的啊？

A：白色、蓝色两款。

在例 2-3 中，Q 与 A 是概念关系匹配，其中颜色与白色、蓝色都是概念关系对。

（3）**句式类型匹配**　　不同句式常有与之对应的固定回答形式；不同类型的引发语有与之对应的应答语。

例 2-4 时间类型的问句，答案通常情况下都与时间相关：

Q：今天什么时候发货啊？

A：都是下午 5 点发货

（4）**公共词序列匹配**　　如果在引发语与应答语中包含相同有意义的词汇序列，表明这两句在论述的内容上有一定相关性。同样，为了保证公共词序列的有效性，统计公共词序列前必需先过滤停用词。

例 2-5 含有公共词序列的问答对：

Q：那我们买了好多东西，可以给我们优惠点吗？

A：买的越多越优惠。

在例 2-5 中，Q 与 A 的公共词序列是“买 多 优惠”。

2.3.2 隐性匹配

隐性匹配是指引发语与应答语两者并没有包含明显匹配的词汇和句法特征，但仍然可以组成有实际意义的相邻语对。判断没有明显特征的两个语句是否匹配通常需要一定的背景知识。

（1）**语义匹配**　　引发语与应答语在语义上是衔接的，但在句法上没有明显的相关特征。仅仅依赖语法特征判断两个语句是否语义匹配，有很大困难。

例 2-6 无明显特征的问答对：

Q：我能抽两次奖吗？奖品有什么呀？

A：说明都在首页有写的。

例 2-6 中，A 不是对 Q 的直接应答，但这两个语句仍然可以作为是问答对。

（2）**以问作答**　　有些情况下引发语中提供的信息不全或有歧义，应答者就会提出疑问，获取更多信息。或者应答语是反问句。

例 2-7 提出疑问，获取更多信息：

Q：我想请问下在店里买 3 样以上可以包邮吗？

A：到那个省哦？

在例 2-7 中，A 是为了获取更多信息，提出疑问。

2.4 问题形式化描述

问句答案匹配要解决的问题是从问题集 Q 中选取一个问句，从答案集 A 中选出一个答案，判断两者是否匹配。可以看成是判断两者是否匹配的二分类问题。

假设，给定一个问题集，其中含有 n 个问句：

$$Q_n=\{q_1,q_2,\cdots,q_n\}$$

给定一个答案集，其中含有 m 个答案：

$$A_m=\{a_1,a_2,\cdots,a_m\}$$

将 Q_n，A_m 两两展开，组成 $n\times m$ 个候选问答对：

$$< q_i, a_j >$$

其中，$i = 1,2,\cdots,n$　　$j = 1,2,\cdots,m$。

给定一个问题q与候选答案a，要判断q和a是否匹配，用函数表述为：

$$f\left(q_i, a_j\right) = \begin{cases} 1 & q_i \text{与} a_j \text{匹配} \\ 0 & q_i \text{与} a_j \text{不匹配} \end{cases} \quad (2\text{-}1)$$

对于问题q有以下几种分类结果：

a) 一个问题q没有与之匹配的答案；

b) 一个问题q有一个与之匹配的答案；

c) 一个问题q有多个与之匹配的答案；

根据检测到与之匹配答案的个数多少，来设定其在知识库的通用性。匹配到对应的数量越多，表明通用性越强。类似的对于答案a也有对应的结果。

2.5 本章小结

本章对问句答案匹配的问题做了详细的描述。首先举例说明了匹配的相关概念，并阐述了其定义。深入分析语料中的语言现象，对其中的问句答案的显性、隐性匹配做详细介绍。最后形式化的描述了问句答案匹配要解决的问题。

第3章 问句答案匹配方法研究

3.1 引言

通过对语料的分析，依据不同的分类思想，设计三种问句答案匹配计算方法，从网络购物的记录中提取有意义的问答对：基于特征匹配的相关度计算方法、基于冗余信息的相关度计算方法、基于词共现的相关度计算方法。下面将详细介绍这个三种计算方法。

3.2 基于特征匹配的相关度计算方法

基于匹配证据的问答匹配的最主要任务是检测问句与答案中是否存在相关的匹配证据。为解决这一问题，本节采用了三种特征检测方法：句式类型、公共词序列、概念关系对。特征的提取是本节整个计算方法中最重要的过程，其效果直接影响到最后的准确率。下面依次对三种特征的提取算法详细说明。

3.2.1 句式类型

想要正确回答一个问句，就要了解问句的需求，即确定问句的类别。依据问句的类别回答搜索答案，不仅可以对回答问句的范围作限制，还能够提出不同的处理策略。例如：“快递到达什么时候呢？”问的是关于时间类别问句。如果系统理解这个问句，对回答问句的范围做出限制，答案的搜索空间将大大减少。

目前并没有标准的问句分类体系，不同的问句分类方法被用于不同的领域。根据本课题语料的特点并参考张耀允等人[39]的三层问句分类，将汉语问句分为两大类：事实类问句、非事实类问句。共 11 小类：5 种事实类问句、6 种非事实类问句、不能识别的问句类型为其他类。如表 3-1 所示：

表 3-1 问句分类体系

大类	小类
事实类问句	实体、地点、人物、数量、时间
非事实类问句	选择、对比、描述、过程、是非、原因

特征选取是问句分类中至关重要的一步，特征选取的效果直接影响问句分类的准确度[40]。文本分类是文本信息处理过程中常用的技术之一，与会话处理过程的问句分类有很大差别。文本分类中包含大量信息，而本文语料中的问句通常是简单句，包含的信息量较少。对于问句分类只选取句子表层的特征，并不能取得理想的分类效果。为此，除了选取问句中的词特征外，还使用了句法和语义特征。

浅层词法特征：词是句子的最基本单位，一个句子由一列词排序组成，且词性是分析句子结构的基础。问句中的疑问词在一定程度上能体现整个问句的需求。因此对每个问题进行分词、词性标注，选取词特征、词性特征作为问句分类的基本特征。

句法特征：在句法分析中，语块是语义表达的基本元素，如命名实体，各句法成份之间的依赖关系等。另外，问句句式模板，如疑问句式，选择句式，并列连词词组等，对问句类别的判断也起很重要的作用。

语义特征：由于一个问句中的词语个数较少，因此通过同义词扩展、HowNet 中上位词对同一类别的词进行泛化。

为了判断哪个特征对问句分类起的作用更大，我们计算了语料中所有特征的信息增益，并中选出最的影响的特征作为分类的特征。一个特征的信息增益值越大，说明该特征对于问句的区分能力越大。

用信息增益的方法选取最有区别度的前 500 个特征，各类别特征的分布如下图 3-2：

表 3-2 前 500 特征中各类别的分布

特征类别	频率
词特征	191
词性特征	164
句法特征	103
语义特征	42

本文采用 SVM 分类方法基于选取的特征进行问句分类。与其它的分类方法相比，SVM 具有较强的泛化能力，能够收敛到全局最优点等特点，因此使用 SVM 算法来进行问句分类。

3.2.2 公共词序列

如果在问句与答案中包含相同有意义的词汇序列，表明这两句在论述的内容上有一定相关性。

本课题的语料环境基于网络购物，在会话的过程中，用户为了表达自己的心情，往往会使用大量的表情符。这些表情符对计算问句与答案匹配的相关度并没有任何贡献，所以在计算相关度前一定要将其过滤掉。表情符都是/:+三个连续的符号或数字组成，如：/:^_^ 、/:）-（ ，可以用正则表达式的方法匹配并过滤。

另外，在网络购物中用户经常在问句中加上表达对其他用户的敬语、称呼及表达自己情绪的词语。如最见的“你好”、“呵呵”、“亲”。这些短语不仅对计算问句与答案的匹配度没有任何帮助，而且会起干扰的反作用。为了更好的体现公共词序列的有效性，首先对句子进行有效性判断、过滤噪音词等预处理。

本文的停用词表是统计现有语料中的会话记录，按照词频倒序排列，经过人工选取并综合信息检索中的停用词表获得。然后用动态规划方法求过滤停用词之后问句与答案的最长公共词序列（LCS）。

例 3-1 含有公共词序列的问答对：

Q：那我们买了好多东西，可以给我们优惠点吗？

A：买的越多越优惠。

在例 3-1 中，Q 与 A 的公共词序列是“买 多 优惠”。

3.2.3 概念关系对

每个概念都有对应的概念集，通过刻画这些对应关系来描述一个概念的性质。匹配的问答对中包含着对应的概念对。

本文的概念关系对主要是基于知网词条的描述获得的。知网中描述了丰富的语义关系。知识词典是知网系统的基础文件，其中的每一个词语的概念及其描述形成一个记录。将每个记录通过一个表述形式关联起来，形成一个知识网络。

知网描述了下列 16 种关系：

表 3-3 知网描述的 16 种关系

（a）相关关系	（b）上下位关系
（c）实体-值关系	（d）同义关系
（e）工具-事件关系	（f）反义关系
（g）场所-事件关系	（h）对义关系
（i）时间-事件关系	（j）部件-整体关系
（k）事件-角色关系	（l）属性-宿主关系
（m）值-属性关系	（n）材料-成品关系
（o）施事/经验者/关系主体-事件关系	（p）受事/内容/领属物等-事件关系

以网购事件为例来说明各概念之间的关系，如图 3-1：

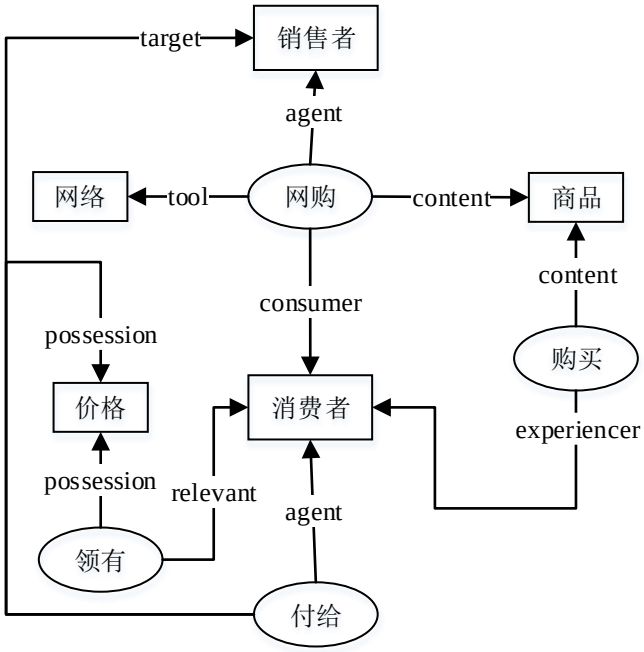

图 3-1 网购事件概念关系图

由于知网词典中的数据都针对通用性的概念设计，并且知网中数据目前没有包含新词，如包邮、快递等。因此本文对语料中词条进行统计，并对知网进行扩展。

在知识词典文件中，一个词条包含词语的编号词性及其描述共 4 项内容。知网的知识词典中关于“红色”的描述如表 3-4 所示：

表 3-4 知网的知识词典关于“红色”描述

No.=35594
W_C=红色
G_C=ADJ
E_C=
W_E=red
G_E=ADJ
E_E=
DEF={aValue|属性值,color|颜色,red|红

在知网中规定 DEF 项中定义的特至少一个，但也可以多个。基于当前词的 DEF 属性查找对应的关系对，如上例中的“红色”与 DEF 属性中的“颜色”。

3.2.4 特征匹配的相关度算法

该算法先是对给定的问句与答案进行预处理，然后检测其中句式类型特征、公共词序列特征、概念关系对特征来构建分类器。根据匹配到特征的个数进行二元分类。具体如算法 3-1 所示：

算法 3-1： 基于特征匹配相关度的算法

输入： 问题q，答案a

输出： 问句与答案匹配输出 1，不匹配输出 0

算法：

1. 问题、答案有效性判断；
2. 对q，a进行分词、词性标注、句法分析、命名实体识别的预处理；
3. 计算问句答案的特征向量

$$v_{ij} =< \text{句式类型，公共词序列，概念关系对} >$$

其中 $i=1,2,\cdots,n;\ j=1,2,\cdots,m$；

4. 依据检测到的特征个数判断给定的问答对否为正例；
5. 将正例保存在问答对集中。

3.3 基于冗余信息的相关度计算方法

冗余信息的相关度的基本思想是正确的信息总会被更多次的重复，而错误的信息则重复度较小。在包含大量信息的知识库中存在大量的冗余信息，因此可以根据信息被重复的次数来判断信息的正确性。

本节是从已有的问答对知识库中，检索出与给定问题相关的答案集，通过计算候选答案与相关的答案集之间的相关度来判断给定问题与候选答案是否匹配。为解决这一问题，首先对问题进行复述，尽可能多的召回知识库中相关问答对，然后从句法结构和语义方面计算两个句子的相关度。最后通过实验结果设定问答匹配阈值。

3.3.1 问题复述

为了更充分的检索出原知识库中的已经存在的相似的问答对，提高召回率，本文采取问句复述的方法对问句进行重构

复述是对句子的另一种表述，在概念上与原句子等价，可以看成相同语义词汇的替换，但目前没有确切的标准表明两个短语或者句子的可替换程度。OrenGlickman 等人认为复述现象体现了自然语言灵活多变性的特性，即一种意思可以用多种形式表达[41]。

复述的主要研究对象是有关短语或简单句的同义表述[42]。据此，复述研究包含两大任务：

a) 构建复述实例语料库，从相关大规模语料中提取表述相近的复述实例；

b) 复述生成技术的研究，利用统计的方法或制定复述规则进行复述。

本文的问题复述主要解决的是同义问题，即对句子中的某个词汇或片段进行同意替换。给定一个问题，进行同义扩展产生一系列相同语义的问题，它们是原问题的复述。例如：“可以给我们优惠点吗？”经过同义词扩展被重写成:“可以给我们便宜点吗？”,“能给我们优惠点吗？”,经过同义扩展，生成一个重构的问句集，目的是尽可能多的找出知识库中具有相同语义的问答对，提高召回率。其中同义词表构建是在哈工大同义词表基础上修改，并加入网络购物中特有的词语，然后将这些重构的问句用问答系统检索，得到一个相关答案集。

本节的主要做法是对问句中的一些词进行同义替换。同义词主要来源于同义词词林哈工大扩展版。《同义词词林》是由梅家驹等人于1983年编纂，

按词义分类编排的类义词典[43]。由于《同义词词林》著作时间久远，其中有很多的词已经不常用。因此哈尔滨工业大学信息检索实验室对其进行了改进，剔除了14706个非常用词，并为了适应当前自然语言处理研究的需要，在词典中加入很多新的词条，形成了一部具有汉语大词表的《哈工大信息检索研究室同义词词林扩展版》，最终的词表包含77,343条词语。在《同义词词林》三级分类的基础上，新的扩展版词典采用五级分类，并将原来的四位编码扩展到八位编码，按照从左到右的顺序排列。第八位的标记有3 种，“=”代表“相等”、“同义”， “#”代表“不等”、“同类”，属于相关词语， “@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。如表3-5所示：

表 3-5 知网 HowNet 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

本文定义两个词的相关度：当两个词的关系为“=”，定义相关度为 2；当两个词的关系为“#”，定义相关度为 1；当两个词的关系为“@”，定义相关度为 0。

原始问题与复述问题的相关性计算，

$$R(Q,Q_i)=\frac{\sum_{i=1}^{k} r(w,w_i)}{k} \tag{3-1}$$

其中， w,w_i 分别表示问题中被替换词、替换词； $r(w,w_i)$ 表示原始问题中被替换的词语与其同义语的相关度，k表示原始问题被替换词的个数。

3.3.2 问答系统检索

当前，辅助问答系统的初始知识库规模为 5 万条问答对，知识库的规模会随着问句答案匹配逐步增加。知识库的规模越大，检索到的相关问答对越匹配。

根据给定的问题检索知识库，并依据问题的相似度返回评分最高的前 5 个。其中问题相关度评分采用 BM25 模型[44]。BM25 算法的主要思想：对给定要检索的查询进行语素解析；计算每个语素与每个搜索结果的相关性得分，

然后，将相关性得分进行加权求和，从而得到查询语句与搜索结果的相关性得分，即每个问答对都有一个基于检索问题的相关度作为权重。

例如：“可以给我们优惠点吗？”的检索结果，如图 3-2 所示：

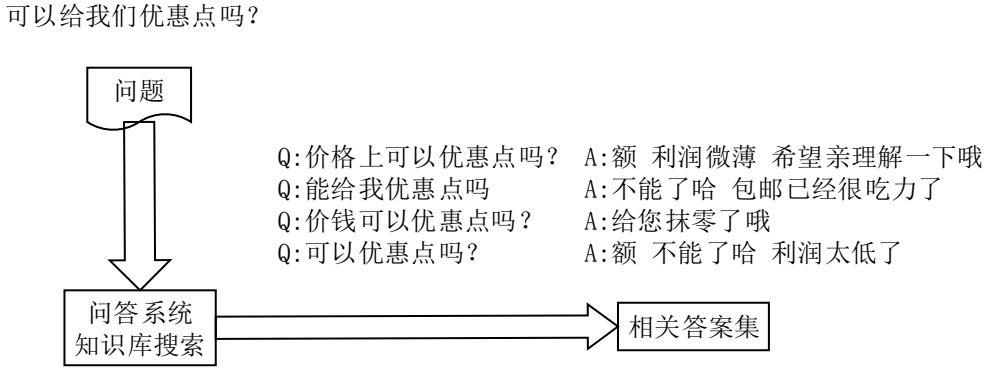

图 3-2 问答系统检索举例

把问题复述后重构的问句都作为检索问句，从现有问答知识库检索相关的问答对，将检索出来的相关问答对及对应的 BM25 相关性评分组成一个有序对，存入问答对集合。计算这些相关的答案与给定答案的相关度，判断给定问句与答案是否匹配。

3.3.3 句子相似度计算

句子相似度是指给定的两个句子之间语意的相关度。计算相关答案与给定答案的相关度，是判断给定问句与答案是否匹配的重要参数。

最常见的两个方法是向量空间模型和编辑距离的方法。向量空间模型的思想：把两个句子都用向量表示，认为每个词组都是独立的作为一个维度，计算两者的夹角余弦。最小编辑距离的思想：指两个字串由一个转成另一个所需的最少编辑操作次数。编辑的操作有三种：插入、删除、替换。这两种方法都不涉及到语义的内容。可能两个句子的词汇很相似，由于结构或出现否定词等造成句义完全不同。由于自然语言的灵活多变，一个句义可以用多种表达方式。因此必须深入到语义的层次才能更好的计算两个句子的相似度。

句子的相似度[45]不仅体现在词汇的相似度，词汇在句子结构中所处的句法成分也起着很重要作用。相同的词汇，不同的组合可能造成完全相反的句义。本文采用李昊迪提出的句法信息与词汇语义的混合方法计算两个句子的相似度。

句子的语义是由词汇和语法结构决定的，因此计算两个句子的相似度必

须兼顾两者。首先采用依存关系的句法结构解析器分析句子成分[26]，选取在两个句子中处于相同结构的词汇，计算它们的相似度，并赋予不同句子成份对应的权值。

利用句法信息与词汇语义的混合方法计算两个句子相似度公式如下：

$$sim_s(S_1,S_2)=\frac{1}{n}\sum_{i=1}^{n}sim_{wMix}(E_{1i},E_{2i})-(n_1+n_2-n)\cdot PF \qquad (3\text{-}2)$$

公式中的句子S_1，S_2分别由n_1,n_2个句法成分组成，n表述公共的句法成分数量，E_1,E_2分别表示S_1,S_2中的每个句法成分。其中E_{1i},E_{2i}表示具有相同的句法功能。对于句法成分不同的短语，不能直接计算它们之间的相似度，除去这些不同成分的句法信息，减少可能造成语义之间的差异。

3.3.4 问句答案匹配度计算

本文的问句答案匹配度主要依据问题与复述问题的相关性、复述问题的检索答案与候选答案的相关性两个指标计算。如下公式：

$$degree(Q,A)=\frac{1}{n}\sum_{i=1}^{n}R(Q,Q_i)\cdot S(A,A_i) \qquad (3\text{-}3)$$

其中，Q表示给定的一个查询，A表示给定的一个候选答案。$R(Q,Q_i)$表示查询Q与问题复述中的一个问题Q_i的相关性。$S(A,A_i)$表示候选答案与复述问题的检索结果的相关性。

复述问题的检索答案与候选答案的相关性计算：

$$S(A,A_i)=\frac{1}{m}\sum_{j=1}^{m}S(A,A_{ij}) \qquad (3\text{-}4)$$

其中，$S(A,A_{ij})$为第i个复述问题检索答案中第j个与候选答案的相关度。m表示第i个复述问题的检索答案个数。

综上，原始问题与候选答案的匹配度计算公式：

$$degree(Q,A)=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{i=1}^{k}r(w,w_i)}{k}\cdot\frac{1}{m}\sum_{j=1}^{m}S(A,A_{ij}) \qquad (3\text{-}5)$$

3.3.5 冗余信息的相关度算法

首先对问题复述，提高问答系统的检索召回率，然后再计算检索答案与候选答案的相关度，最后计算原始问题与候选答案的相关度。基于冗余信息的相关度计算方法流程如图 3-3 所示：

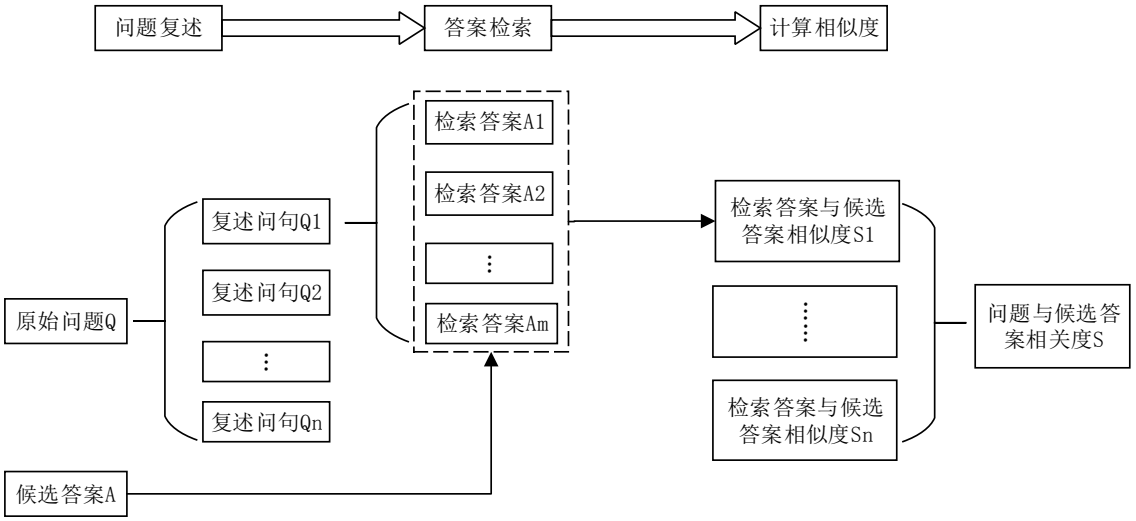

图 3-3 冗余信息的相关度计算方法流程图

3.4 基于词共现的相关度计算方法

词共现是指在一些词项经常在同一文档中出现。词共现类似于词语搭配，但又不仅限于搭配，共现词可以是习惯搭配关系的词对；也可以是属于相同词义的词对，例如快递与物流、优惠与便宜等；或者是在同一话题中经常出现的词对，例如：经常与包邮同时出现的词有地点、时间、价格等。

词共现计算方法是一种统计方法。它是建立在这样一个基本假设的基础之上：如果在大规模语料中，两个词经常共现在文档的同一窗口单元（如一句话、一个自然段等），则认为这两个词在意义上是相关的，并且共现的概率越高，其相关度越紧密。

3.4.1 词共现相关度训练策略

预处理在该计算方法中尤为重要，因为在整个问答对知识库中出现频率比较高的高频词往往是那些无实际意义的词，如：“的”，“哈哈”，“亲”。在统计词共现频率的时候，这些词与其他有意义的词一起出现的频率也会变的

很高，然而这些无实际意义的词组成的高频词对到计算问句与答案的相关性并没有重要任何帮助。因此有必有对问句与答案做预处理。类似于公共词序列，词共现计算方法的问句同样需要预处理：句子有效性判断、不规则符号过滤、分词、噪音词过滤。

通常从两个角度分析两个词汇的相关度[46]：第一，计算两个词同时出现在一个窗口单元的相关度；第二，计算两个词的共现窗口在一篇文档中多次出现的相关度。不能把两个问题孤立起来，根据不同的应用环境，叠加成相应的复合模型。本文中词共现是指，问答对中经常对应出现的词项，将问答对作为一个共现窗口，统计整个知识库中词项对出现的频率，作为词对的相关度。

本文词共现的窗口概念与文档词共现统计的窗口概念不同。在文档统计中，一个共现窗口通常指一个语句、一个段落或者一篇文章，窗口内的所有词都可以组成共现对。在本文中，窗口是一个问答对，并且只有问题中的词条与答案中的词条才能组一个共现对。

问句与答案都是简单句，句子长度有限，因此以问答对为一个共现窗口，不存在普通词共现模型中第二种情况，即在一个文档中有多个词汇共现窗口单元。

本文采用简单的词共现统计计算方法，统计整个知识库中词项对出现的频率，作为词对的相关度。

$$R_D([a,b]) = G(a,b,f(a,b)) \qquad (3\text{-}6)$$

其中 $f(a,b)$ 为 a 和 b 共同出现的频率。

为了最大的限度的减少无区分度词的干扰，除了对问题、答案做预处理，还采取“慢加快减”策略，即正例词对频率每次加 1，反例词对频率在原基础上减半。如果标注的语料是正例，对问句与答案中共同出现的词语对的频率加 1；如果标注的语料是反倒，则将问句与答案中共同出现词语对的频率减半，即把原词语对的频率除以 2。

3.4.2 词共现的相关度算法

统计相关语料中经常与目标词共同出现的词汇的频率，用频率较高的词汇构成高频词对。统计标注后的训练语料中问句答案匹配出现的关联词汇，并形成关联矩阵，表述如下：

$$\begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{array} \begin{array}{c} \begin{array}{cccc} b_1 & b_2 & b_3 & b_m \end{array} \\ \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{c2} & \cdots & c_{nm} \end{pmatrix} \end{array} \tag{3-7}$$

其中 $a_1,a_2,\cdots,a_n$ 表示问题中出现的词汇集，$b_1,b_2,\cdots,b_m$ 表示答案中出现的词汇集，c_{ij}表示a_i与b_j匹配的相关度。本文以a_i与b_j共同出现的词汇频率作为相关度。

问题 $q=<a_{i1},a_{i2},\cdots,a_{is}>$ 与答案 $a=<b_{j1},b_{j2},\cdots,b_{jt}>$ 都可看作为词汇向量。首先，提取问句答案匹配的关联矩阵 $\boldsymbol{A}$：

$$A=\begin{array}{c} \\ a_{i1} \\ a_{i2} \\ \vdots \\ a_{is} \end{array} \begin{array}{c} \begin{array}{cccc} b_{j1} & b_{j2} & b_{j3} & b_{jt} \end{array} \\ \begin{pmatrix} c'_{11} & c'_{12} & \cdots & c'_{1t} \\ c'_{21} & c'_{22} & \cdots & c'_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ c'_{s1} & c'_{s2} & \cdots & c'_{st} \end{pmatrix} \end{array} \tag{3-8}$$

问题中一个词与答案相关度d定义为：

$$d=\frac{\left\|a_{ij}\right\|}{n} \tag{3-9}$$

其中，$\left\|a_{ij}\right\|$ 为矩阵$\boldsymbol{A}$中第 i 行向量的模长，n 为答案中包含词的个数。例如：a_{i1}与答案 $a=<b_{j1},b_{j2},\cdots,b_{jt}>$ 的相关度为 $d_{a_{i1}}$：

$$d_{a_{i1}}=\frac{1}{t}\sqrt{c'^{\,2}_{11}+c'^{\,2}_{12}+\cdots+c'^{\,2}_{1t}} \tag{3-10}$$

问句与答案相关度D定义为：

$$D=\frac{\sum_{i=1}^{n} q_i}{n} \tag{3-11}$$

其中，q_i为问句中第i个词。

具体计算过程如算法 3-2 所示：

算法 3-2：基于词共现相关度算法

输入： 问题q，答案a

输出： 问句与答案匹配输出 1，不匹配输出 0

算法：

1. 问题、答案有效性判断；
2. 对q，a进行分词、去停用词的预处理；
3. 计算问题中的每个词与答案的相关度d；
4. 依据第 3 步结果计算给定问句与答案的相关度D；
5. 将最后评分高于阈值的正例保存在问答对集中。

3.5 本章小结

本章主要研究了问句答案匹配的三个计算方法，详细介绍用到的相关算法及流程。首先说明特征匹配的计算方法，然后阐述基于冗余信息的相关度计算方法，接着介绍了词共现计算方法。每个方法都有自己的特点。第一个方法是基于特征检测的算法，其它两个都是基于统计的算法。基于特征的算法，不能够覆盖到所有的特征词，但是准确率稳定。基于统计的算法，太过依赖于语料。三个计算方法相互补充。

第4章　语料标注及分析

4.1 语料基本信息统计

本文的研究工作以网上购物的真实聊天数据为基础，收集的语料涉及 4 类商品，一共 38069 组会话，随机从中选出 174 组会话，5827 个语句，平均每组对话 33 个语句。食品 57 组、手机 41 组、玩具 31 组、衣服 45 组。

首先对语料中的会话平均长度，句子长度基本信息进行统计和分析，统计结果如表 4-1：

表 4-1 网上购物语料基本信息表

	每个会话长度（话轮）			问句话轮中语句个数			回答话轮中语句个数			每个语句的长度（字数）		
	AVG	MIN	MAX	AVG	MIN	MAX	AVG	MIN	MAX	AVG	MIN	MAX
食品	30	16	35	1.26	1	3	1.20	1	4	6.15	1	92
玩具	35.14	16	52	1.38	1	3	2.13	1	8	5.20	1	77
手机	31.66	10	46	1.63	1	5	2.03	1	6	4.88	1	117
衣服	36.46	18	44	1.40	1	4	1.47	1	6	10.67	1	169

会话长度以问答的轮数计算。本文收集的语料中最短的只有 10 轮，最长的达到 52 轮，平均每个会话都达到 20 轮以上。

从问句话轮中语句个数可以看出一个话轮中通常含有多个语句，最多的达到 8 个。回答话轮中语句个数多于问句话轮中语句个数。

每个语句的长度以字数计算。在交互式会话过程中，一般习惯用 10 字以内的简单语句表述。不同于张耀允等人[39]用 wizard-of-oz 方法收集的语料统计，她们认为交互问答中复杂问句比例占 2/3 以上。

基于上面的统计，一个会话中包含多轮话轮转换，在每个话轮中包含多个简单语句。

4.2 非规范语言现象的统计

观察发现语料中存在大量的非规范语言表述，有的达到滥用的地步。本文研究归纳总结了以下几种语料中存在的非规范语言现象：

（1）**口语词**　　口语灵活多变，因场合与说话人不同而被自由使用。

在聊天过程中，口语的简洁性、自由性，不仅不妨碍理解，而且更形象地体现说话人的态度。但对于自然语言的理解则是非常大的干扰。因为其灵活性，口语是出现频率最高的非规范语言。

例 4-1 几个常用口语词：

a) 不能的哈　　b) 咱来买东西咯　　c) 好的哦

（2）**省略**　　省略是指句子缺少应该具备的语法成分。虽然省去句子语法构造必要的组成部分，但根据上下文仍能表达出完整的意思。句子成份的省略使分词、句法分析的准确率下降，并且由于语义不完整给信息检索、语义分析造成很大困难。

例 4-2 下面两个句子句法成份不完整，但在实际交流中不影响理解：

a) 包邮？　　b) 很薄的

（3）**表情符及网络语**　　网络语有多种类型是一种不同传统的语言形式，它流利于网络上，并随着网络的发展不停的丰富和淘汰。表情符和语气词一样，可以生动地体现说话人的心态。

例 4-3 常用网络语举例：

a) /:（ok）　　b) 亲，欢迎光临

上述 3 种非规范语言现象出现的比例统计如表 4-2：

表 4-2 非规范语言现象出现比例

	口语词	省略	表情符及网络语	多现象并存
食品	27.06%	21.83%	19.01%	9.25%
玩具	26.91%	22.34%	18.46%	8.72%
手机	28.58%	21.67%	20.01%	10.15%
衣服	27.92%	21.76%	20.48%	9.94%

出现比例以包含非规范语言现象的对话子句中在一组对话中占的百分比计算。多现象并存是指一个对话子句中至少包含两种非规范语言现象。

基于对语料的统计，语气词的使用频率最高。语气词、表情符及网络语等无实际语义的词汇出现比例近 50%。在非规范语言上，网上购物的交互式对话语言特点明显的不同于人与人直接口语交流的对话特点。宗成庆等人[47]统计的汉语口语对话语料中重复、冗余、次序错位现象出现比例近 10%，而本文收集的语料中基本上不存在这三种现象。

本文收集的语料中除了上述大量的非规范语言现象外，还有错别字、代

词“他”、“它”，“哪个”、“那个”混淆。

提高系统对语句的理解，必须能够识别并预处理非规范语言现象。

4.3 问句答案匹配现象的统计

根据第 2 章中问句答案匹配现象的分析，对收集的语料进行统计，其中显性匹配主要指通过问句与答案中一些词的特征现象来匹配，隐性匹配主要指语义级别的匹配。统计结果如表 4-3 所示：

表 4-3 问答对匹配现象出现比例

	显性匹配				隐性匹配	
	指代匹配	概念关系	句式类型	公共词序列	语义匹配	以问作答
食品	13.72%	13.12%	37.72%	18.17%	6.85%	5.66%
玩具	15.33%	10.24%	30.31%	15.61%	7.53%	4.12%
手机	10.64%	9.33%	32.89%	12.26%	7.72%	5.91%
衣服	14.81%	12.42%	38.74%	20.31%	6.61%	3.79%

在统计问句相关现象时，如果简单的应答语句只含有：哦，嗯，表情符等无意义词不作为相邻语对匹配情况处理。表中的数据是一个会话中各个特征所占的平均比例。每个会话中相邻语对匹配所有特征比例和为 75%以上。

从表中的统计数据来看，约 60%的相邻语对为显性匹配相关，其中句式匹配特征出现比例最高。隐性匹配关系中，语义匹配和以问作答出现比例相当。

基于上面相邻语对匹配的统计，显性匹配所占的比例远大于隐性匹配。对无明显匹配特征的隐性匹配的识别与检测，特别是对语义匹配的检测有存在一定困难。

4.4 语料标注及有效问答对统计

为了测试第 3 章中设计的计算方法，需要收集和标注一些语料作为训练集和测试集。

（1）训练集的收集。所有语料涉及 4 类商品，一共 38069 组会话，一共 103 万多条会话记录。在这些记录中存在着一少部分一一对应的简单语境下的问答对，共 2 万条，这些问答对基本上都是匹配的。因此可以把这 2 万条问答对作为训练集。

（2）测试集的标注。从3万多组会话中选取100组会话进行人工标注：标注方法采取每组会话3人标注，对标注结果不一样的取多数标注结果。将标注好的结果作为测试集，验证计算方法的性能。

（3）有效问答对统计。在会话记录中包含很多简单的问候语或应答语，如：“是的”、“亲”等，这些句子并没有一定的语义，与其它句子组合成的问答对也不具有可用性。因此在判断匹配前，有必要把这些无意义的句子过滤掉。

由于网络购物会话中的问句与答案是交叉匹配的，无法根据所处的相对位置判断两个句子是否匹配，所以有必要把一个话轮中的问句与答案的对应关系展开，两两组合。

例4-4 下面的话轮中，用户连续问两个问题，客服分别回答：

Q1：如果不发申通，还能发什么快递

Q2：你能不能给我包邮啊

A1：我们的默认为申通及韵达快递

A2：可以包邮的，亲

把这个话轮两两展开，组合四个问答对，Q1与A1，Q1与A2，Q2与A1，Q2与A2。而实际问句与答案匹配的问答对只有：Q1与A1， Q2与A2。

有效问答对统计包含两个步骤：

a) 有效句子统计

经统计发现，网络购物语料中的语句，很多都是简单的回应语，如：“好的”，“嗯”，“是的”，“没有”等。这些句子并没实际的意义，因此有必要对句子的有效性做出判断。对句子进行特征符号，噪音词过滤，如果句子长度大于2，则作为有效的句子。

b) 问句与答案匹配统计

把一个话轮中的问句与答案两两展开，对展开的问答对进行人工标注，判断是否匹配。不同商品的会话中的有效问答对比例都比较偏低，有效句子的平均比例约为35%。匹配的问答对在有效句子的占的比例约为37%。有效问答对比例统计如图4-1：

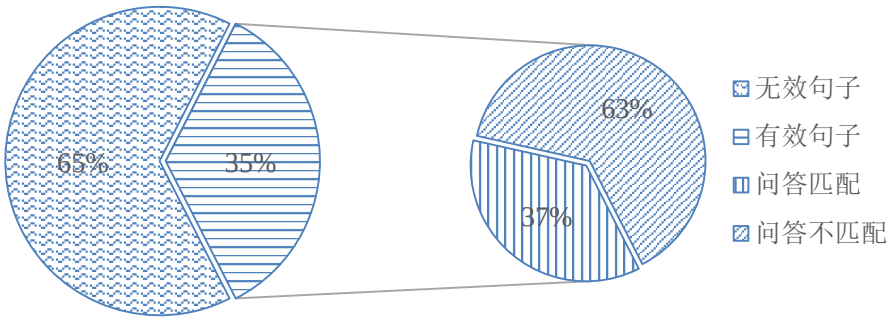

图 4-1 有效问答对比例统计图

4.5 本章小结

本章主要对收集到的语料做系统的统计分析，除了统计语料的基本信息，还对语料进行了标注，统计其中的语言现象。介绍了实验的训练集和测试的收集标注方法，为第 5 章实验做准备。由于语料是基于网络的会话记录，在语料中包含大量干扰、无意义的语句，因此本章还对会话中有效句子比例进行了统计。

第5章 问句答案匹配系统实现与评测

5.1 引言

设置不同的实验分别测试三个计算方法从网络购物语料中提取问答对的效果，并对比三个计算方法的实验结果，分析各个它们的特点。最后将三个计算方法有机结合成一个自训练模型框架，测试其实验效果。

5.2 评价标准

本文实验结果的分析，评价标准主要有：

（1） 准确率

准确率指提取出的问答对中正确的比例，公式表示为：

$$P=\frac{\text{正确的问答对数}}{\text{提取出的问答对总数}} \tag{5-1}$$

（2） 召回率

召回率指提取出的问答对占整个对话中问答对的比例，公式表示为：

$$C=\frac{\text{提取出的问答对数}}{\text{一组对话问答对总数}} \tag{5-2}$$

（3） F 值

F 值是综合准确率与召回率对进行整体评测公式如下：

$$F_{\beta}=(1+\beta^{2})\cdot\frac{P\cdot R}{\beta^{2}\cdot P+R} \tag{5-3}$$

由于提取出的问答对将作为正例存入知识库，所以准确率比召回率重要。所以计算 F 值的时候将准确率的权重提高，取$\beta=0.5$。

5.3 三个计算方法实验结果对比及分析

（1）基于特征匹配的相关度计算方法实验结果

该计算方法主要采用三种特征来检测问句与答案的匹配证据。当寻找到的匹配证据个数为多数或都有效的公共词序列长度大于 3 时，将相应的问答对作为正例。准确率与召回率是在问句与答案都是有效句子的基础上计算的。实验结果如下表 5-1 所示：

表 5-1 基于特征匹配的相关度计算方法实验结果比较

	准确率	召回率	$F_{0.5}$ 值
特征匹配结果	0.75	0.65	0.73
BaseLine 结果	0.35	0.50	0.37

注：BaseLine 是以随机分类方法的实验结果。

从 BaseLine 中可以看出，一个会话中匹配的问答对比较少。通过特征检测方法来寻找问答匹配的证据，能够有效的提取出问答对。

（2）基于冗余信息的相关度计算方法实验结果

该计算方法首先通过问题复述构建多个问句，检测现有知识库，以提高问答系统的召回率，然后计算检索到答案与候选答案的相关度。最后得出给定的问题与候选答案的相关度。

分别选取 2 万条问答对集和 5 万条问答对集作为现有系统的知识库，分别实验，并对比实验结果。实验统计结果如图 5-1 所示：

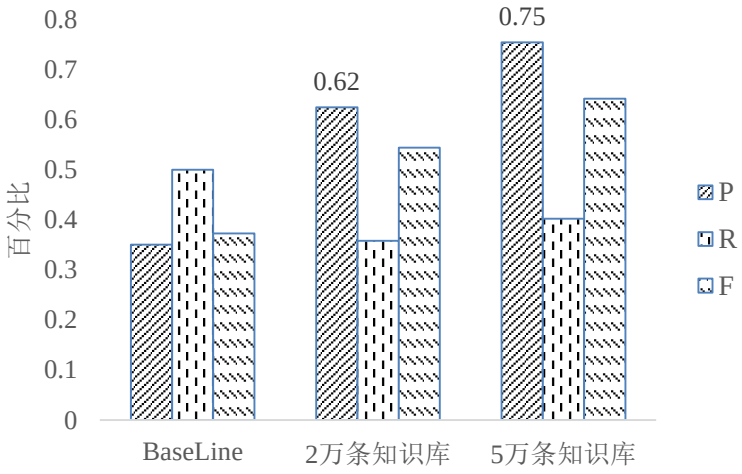

图 5-1 基于冗余信息的相关度计算方法实验结果对比

首先从包含 2 万条问答对的问答系统中检索，并计算问题与候选答案的相关度。然后将问答系统的知识库扩充到 5 万条问答对，再次实验。从实验结果中可以看出，基于冗余信息相关度计算方法能够有效的从网络购物语料中提取出问答对，并且随着问答系统知识库的规模扩大，效果也会明显的提高。

（3）基于词共现的相关度计算方法实验结果

本章通过设置不同实验训练集，分别实验，并对比实验结果。首先将标注语料中的 1 万条问答对作为训练集，根据算法 4-2 训练词共现计算方法。然后将训练语料扩大到 1.5 万条，在相同的测试集上实验。最后将所有标注语料都用到训练，并在测试集上实验。三次实验结果如 5-2 所示：

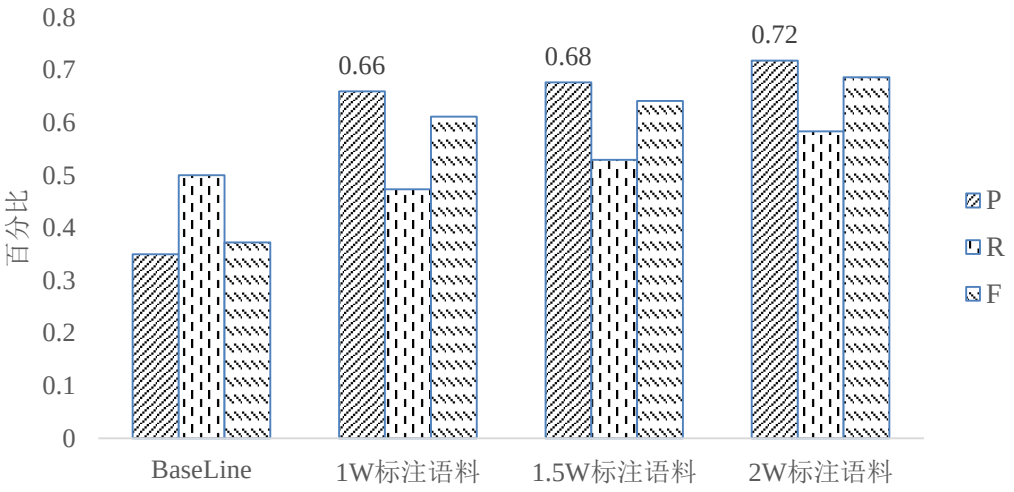

图 5-2 基于词共现的相关度计算方法实验结果

从实验结果可以看出，词共现的相关度计算方法能够有效的提取出匹配的问答对，但是随着训练语料的扩大实验效果并没有明显的提高。表明该计算方法很大的瓶颈。

（4）三种计算方法结果比较

三个计算方法的目的是从网络购物语料中提取有意义的问答对，并将这些问答对放入到知识库中。由于准确度是一个最重要的衡量标准，先对三个计算方法的准确度及其波动性比较如图 5-3 所示：

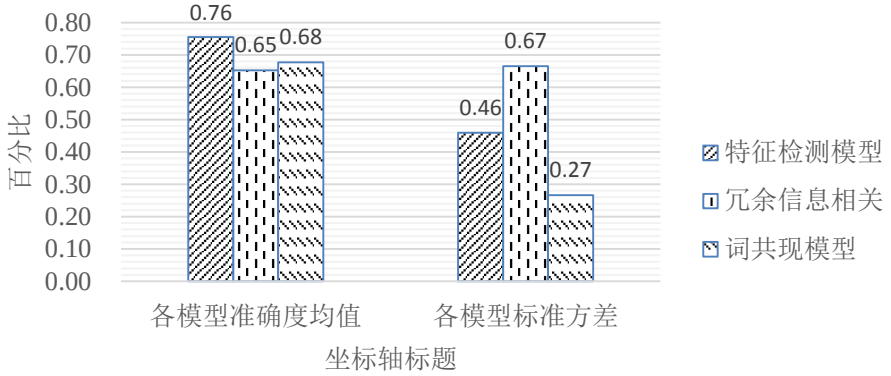

图 5-3 三个计算方法的准确度及其方差比较

从对比结果中可以看出，基本特征匹配的计算方法的准确度比较高，基

于词共现的相关度计算方法的稳定性最高。三个计算方法中效果最差的是基于冗余信息的相关度计算方法，不仅准确度不高，而且波动率比较大。但是在冗余信息计算方法的实验中，随着知识库的增加，该计算方法性能有明显的改善。因此在以后的实际应用中，其性能将会大提高。

5.4 构建自训练模型框架

由 5.3 节的实验结果可以看出三个模型各有不同的特点：特征匹配方法虽然准确率得到保证，但是无法覆盖全部特征，并且不具有自动学习的能力；冗余信息方法不仅准确率低，而且波动性最大，无法保证提取问答对的有效性，但是随着知识库的扩大算法的性能有显著的改善；词共现方法有比较稳定的性能，但是随着训练语料的扩大整体性能提高有限。

整个框架流程如图 5-4 所示：

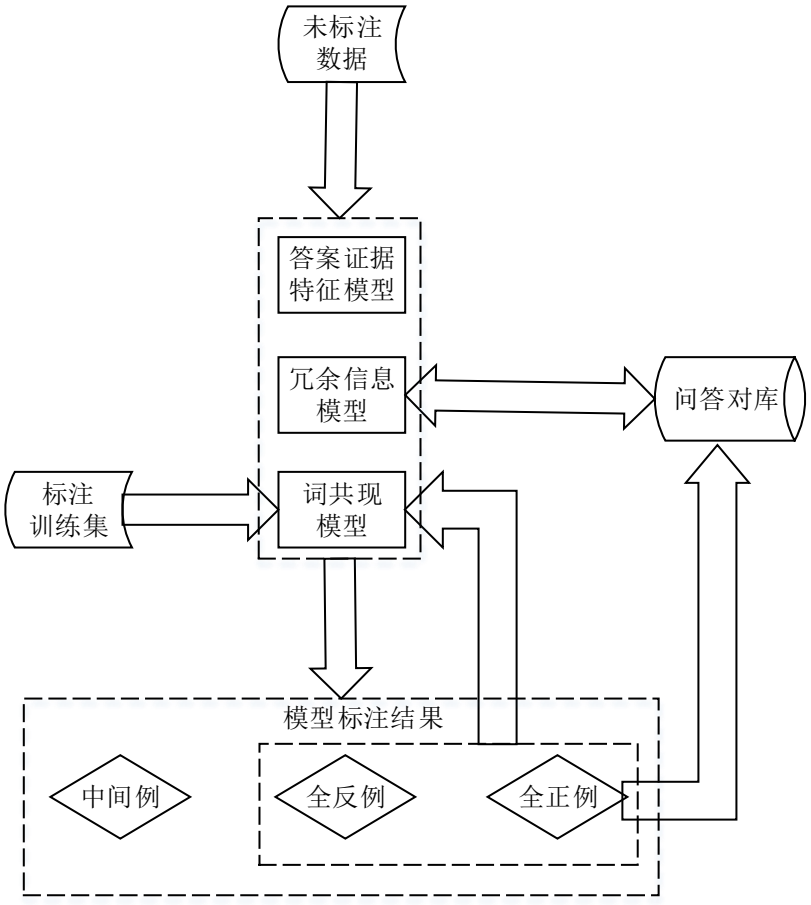

图 5-4 问句答案匹配系统流程图

本文构建的自训练框架中三个计算方法相互监督，先用少量标注的语料

训练算法，利用训练过的初始算法对未标注的语料分类，将分类出来的结果作为训练集，对算法进行再次训练。类似的进行重复迭代直到整个框架性能稳定。

传统的基于特征检测的方法有一个致命的缺点，就是无法枚举所有可能的关键词、句式等线索特征。传统的标注统计的方法有一个明显的不足，就是为了训练一个分类模型需要标注大量的语料。传统的自训练方法通过迭代自己标注的实例来训练模型。这种迭代过程虽然可能会提高置信度，但是并不意味着分类的结果没有错误实例。错误分类的实例将会导致下次更坏的迭代结果。只有当所有分类结果全为正例的实例才能进入下次迭代。这样就大大降低了错误的积累，将反例的影响限制到最小。

本文设计的自训练模型框架对问句与答案是否匹配进行检测在一定程度上解决传统方法中两大挑战不足。该框架可以用少量标注的数据集对大量的未标注数据进行迭代分类。

5.5 自训练框架实验结果分析

自训练框架的实验结果取决于三种计算方法及判断是否匹配的策略。当三种方法的性能都提高时，自训练框架的性能也随之提高。匹配策略有两种：第一种是使用投票策略，当三种计算方法的结果有两个为正时，判断问句与答案匹配；第二种是严格的策略，即当三个结果全为正例是才判断问句与答案匹配。将三个方法结果都为反例的时候，判断问句与答案不匹配。

对三个初始计算方法进行测试，全为正例及有两个正例的统计结果如图 5-5 所示：

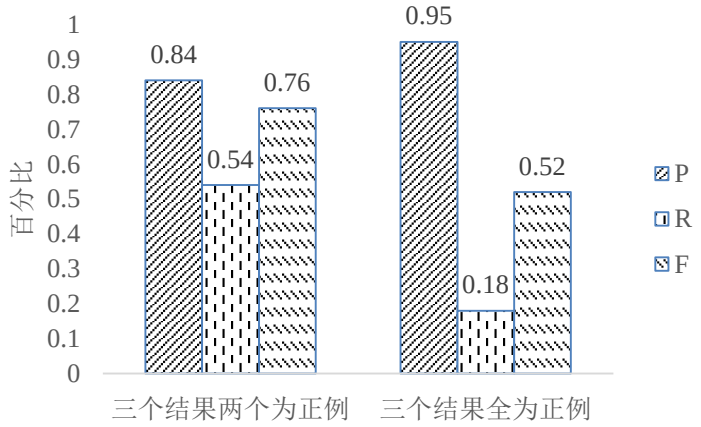

图 5-5 三个计算方法结果全为正例和有两个正例的统计结果

当三个计算方法中有两个为正例时，综合实验结果相比较单一方法的准

确率大大提高，约提高 10%，并且召回率没有明显的下降。当三个结果全为正例的统计结果表明，准确率为 95%，然而召回率却急剧下降，仅有 18%左右。为了保证一定的召回率，在自训练框架中采取投票策略。

自训练模型第一次迭代训练前从未标注的数据集中共提取出 33518 个结果。结果分布如表 5-2 所示：

表 5-2 自训练模型从未标注数据集中提取结果

模型分类结果	比例
都为正例	8.77%
两个正例	19.00%
一个正例	37.02%
全为反例	35.21%

根据投票策略判断分类结果，选取其中匹配的问答对作为正例，不匹配的问答对作为反例，迭代训练模型。

类似的每当提取出的问答对正例增加 5000 对，进行一次训练。经过多次训练整个模型的性能如图 5-6 所示：

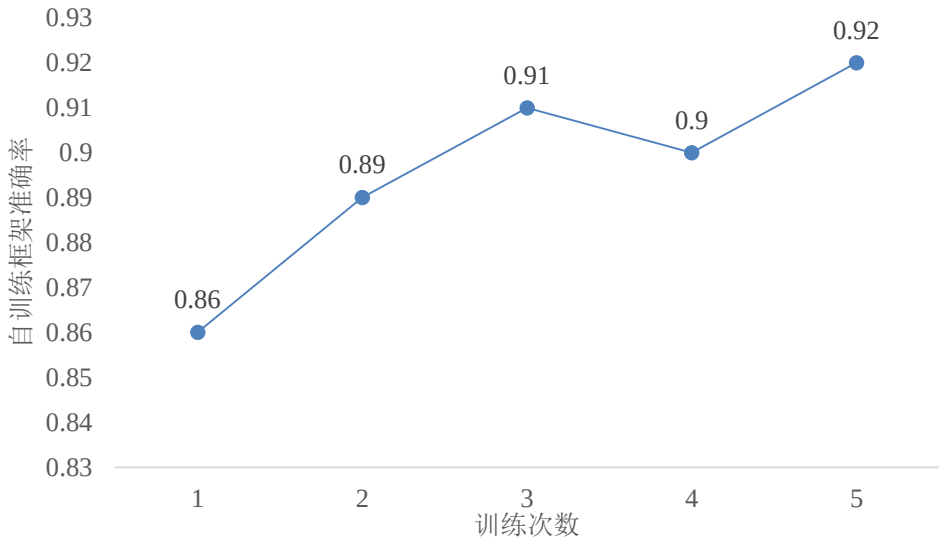

图 5-6 自训练框架多次训练后准确率对比

从图 5-6 中可以看出，经过前面三次训练模型的准确率有显著的提高。

随着训练次数的增加模型性能虽然有一定的波动，但整体趋势向着好的方向逐渐改善。

模型收敛的定量分析：特征匹配的计算方法不具有自动学习功能，其准确率不变。冗余信息的计算方法和词共现的计算方法起初因为训练语料不充分导致准确率偏低，随着训练次数的增加，将提取到的问答对加入到训练集中，它们的准确率逐渐提高。最后当三个计算方法的准确率都达到上限时，整个模型的准确率趋于稳定。在训练的过程中由于反例的影响会造成冗余信息的计算方法和词共现的计算方法准确率的波动，但随着训练集的增加，反例造成的错误会被慢慢纠正。

5.6 本章小结

本章介绍了实验的设置及评价标准，并对比、分析了三种计算方法的实验结果。实验结果表明，每个计算方法都能有效的从网络购物记录中提取问答对，基于特征匹配的计算方法优于另外两个。将它们三个有机组合成自训练模型框架，大大提高准确率。

结　论

网络购物语料中存在多问句与多答案交叉的复杂对应关系，其最大特点就是答案的滞后性。为了解决从复杂对应的记录中自动提取问答对的问题，本课题收集了大量会话记录并进行标注与分析，设计了三种计算问句答案匹配的方法：

（1） 基于特征匹配的相关度计算方法是用特征提取的方法从给定的问句与答案中的检测特征，并依据提取到的匹配特征数判断两者的相关性。

（2） 基于冗余信息的相关度计算方法是从现有的知识库中检索给定的问题，得到相关答案集，计算相关答案与候选答案的相关度来衡量给定问题与候选答案的相关度。

（3） 基于词共现的相关度计算方法是以问答对为共现窗口，统计标注语料中的共现词出现的频率作为共现词对的相关度，然后基于词的相关度计算两个句子的相关度。

通过实验比较三种计算方法的优势与不足，将它们有机组合起来，设计一个自动从大量的网络购物记录中提取问答对的自训练模型框架。该框架能够基于小规模的标注语料，迭代分类大量未标注的语料，大大降低人工标注的劳动量，其系统性能显著高于三个单一的计算方法。

本文设计的自训练问句答案匹配模型框架为从网络购物会话记录中提取出有效的问答对及问答知识库的建设提供了新的思路及解决方案。但是整个系统的执行效率及提取问答对的准确率仍有进一步的提高空间。

自训练模型框架的性能依赖于三个计算方法的效果，每个计算方法的性能将会最终结果产生直接影响，将来改进方向是提高每个计算方法的效果：

（1） 对于基于特征匹配的计算方法只用到三个特征，并且没加入本体知识库和文本推理机制致使该计算方法的召回率比较低。在将来的工作中提取更多的特征，提高计算方法的检测性能。

（2） 冗余信息计算方法依赖于问答系统的检索和知识库的规模和质量，目前系统检索策略只是计算问题之间的相似度，并且稳定性不够。当前系统的知识库仍在初期阶段，规模比较小，需要大幅增加知识库的规模。

（3） 词共现计算方法中仅仅简单的将词对共同出现的频率作为两个词的相关度，并没有考虑到这些词在句子中的句法对应关系。将来改进方向考虑加入句法分析，更加准确的计算两个句子的匹配度。

参考文献

[1] Mollá D, Vicedo J L. Question answering in restricted domains: An overview[J]. Computational Linguistics, 2007, 33(1)：41-61.

[2] Harabagiu S, Moldovan D, Clark C, et al. Answer mining by combining extraction techniques with abductive reasoning[C]//Proceedings of TREC. 2003, 2003：375-382.

[3] 陈振宇，袁毓林，张秀松，等. 一种基于大知识库的亲属关系自动推理模型[J]. 中文信息学报，2010，24(3)：51-53.

[4] Green Jr B F, Wolf A K, Chomsky C, et al. Baseball: an automatic question-answerer[C]//Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. ACM, 1961：219-224.

[5] Zheng Z. AnswerBus question answering system[C]//Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002：399-404.

[6] Kupiec J. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia[C]//Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1993：181-190.

[7] Hammond K, Burke R, Martin C, et al. FAQ finder: a case-based approach to knowledge navigation[C]//Artificial Intelligence for Applications, 1995. Proceedings., 11th Conference on. IEEE, 1995：80-86.

[8] Buckley C, Salton G, Allan J, et al. Automatic query expansion using SMART: TREC 3[J]. NIST SPECIAL PUBLICATION SP, 1995：69-69.

[9] Garofolo J S, Auzanne C G P, Voorhees E M. The TREC spoken document retrieval track: A success story[J]. NIST SPECIAL PUBLICATION SP, 2000(246)：107-130.

[10] Buckley C, Singhal A, Mitra M, et al. New retrieval approaches using SMART: TREC 4[C]//Proceedings of the Fourth Text REtrieval Conference. 1995：25-48.

[11] Aliod D M. Answerfinder in TREC 2003[C]// Proceedings of TREC. 2003：

392-398.

[12] Voorhees E M. The TREC 2004 robust retrieval track[C]//Proceedings of TREC 2004. 2004：102-104.

[13] Fukumoto J, Kato T, Masui F. Question answering challenge (qac-1) an evaluation of question answering task at ntcir workshop 3[C]//Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering. National Institute of Informatics, 2003：1-3.

[14] Kishida K, Chen K H, Lee S, et al. Overview of CLIR task at the fourth NTCIR workshop[C]//Proceedings of NTCIR. 2004, 4：1-38.

[15] 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, 2005：5-7.

[16] Sakai T, Kando N, Lin C J, et al. Overview of the ntcir-7 aclia ir4qa task[C]//Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, Tokyo, Japan. 2008：100-110.

[17] Magnini B, Romagnoli S, Vallin A, et al. The multiple language question answering track at clef 2003[C]//Comparative Evaluation of Multilingual Information Access Systems. Springer, 2004：471-486.

[18] Laurent D, Séguéla P, Nègre S. Cross lingual question answering using qristal for clef 2006[M]//Evaluation of Multilingual and Multi-modal Information Retrieval. Springer Berlin Heidelberg, 2007：339-350.

[19] Gey F, Larson R, Sanderson M, et al. GeoCLEF 2006: the CLEF 2006 cross-language geographic information retrieval track overview[M]//Evaluation of Multilingual and Multi-modal Information Retrieval. Springer Berlin Heidelberg, 2007：852-876.

[20] Forner P, Peñas A, Agirre E, et al. Overview of the clef 2008 multilingual question answering track[C]//Evaluating Systems for Multilingual and Multimodal Information Access. Springer, 2009：262-295.

[21] Van Schooten B, Op Den Akker R. Follow-up utterances in QA dialogue[J].

TAL. Traitement automatique des langues, 2005,46(3)：181-206.

[22] Ferrucci D A. Introduction to “this is watson”[J]. IBM Journal of Research and Development, 2012(56)：1.

[23] Lally A, Fodor P. Natural Language Processing With Prolog in the IBM Watson System[J]. Retrieved June, 2011：15-21.

[24] Kaistinen M, Ryhänen P. Concurrent and Parallel Computing IBM Watson project[J]. 2011：10-15.

[25] Carenini G, Murray G. Visual structured summaries of human conversations[C]//Proceedings of the first international workshop on Intelligent visual interfaces for text analysis. ACM, 2010：37-40.

[26] 王树西. 问答系统：核心技术，发展趋势[J]. 计算机工程与应用，2005,41(18)：1-3.

[27] 毛先领，李晓明. 问答系统研究综述[J]. 计算机科学与探索，2012,6(3)：193-207.

[28] Raine R. Making a clever intelligent agent: The theory behind the implementation[C]//Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on. IEEE, 2009, 3：398-402.

[29] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 中国科学院计算机技术研究所博士学位论文，2004：12-14.

[30] 王树西，白硕，姜吉发. 基于自由文本的模式推理[C]//见：第一届全国信息检索与内容安全学术会议. 2004：354.

[31] 张巍，陈俊杰. 浅层语义分析及 SPARQL 在问答系统中的应用[J]. Computer Engineering and Applications, 2011(47)：65-69.

[32] Huang J, Zhou M, Yang D. Extracting chatbot knowledge from online discussion forums[C]//Proceedings of the 20th international joint conference on Artifical intelligence. Morgan Kaufmann Publishers Inc., 2007：423-428.

[33] 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.

[34] 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.

[35] Ding S, Cong G, Lin C, et al. Using conditional random fields to extract contexts and answers of questions from online forums[J]. 2008：10-21.

[36] 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, 2011,47(6)：886-898.

[37] Wang B, Liu B, Sun C, et al. Extracting Chinese question-answer pairs from online forums[C]//Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. IEEE, 2009：1159-1164.

[38] Sacks H, Schegloff E A, Jefferson G. A simplest systematics for the organization of turn-taking for conversation[J]. Language, 1974：696-735.

[39] 张耀允，王晓龙，王轩，等. 面向开放的限定领域的交互式问答语料分析[J]. 中国计算语言学研究前沿进展, 2011：235-243.

[40] Leuski A, Patel R, Traum D, et al. Building effective question answering characters[C]//Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics, 2009：18-27.

[41] Dagan I, Glickman O, Magnini B. The pascal recognising textual entailment challenge[M]//Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. Springer Berlin Heidelberg, 2006：177-190.

[42] 赵艳. 基于相关短语挖掘的问句复述研究[D]. 哈尔滨工业大学, 2009：45-48

[43] 梅家驹. 同义詞詞林[M]. 上海辞书出版社, 1983：1-5.

[44] Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004：42-49.

[45] Mohler M, Mihalcea R. Text-to-text semantic similarity for automatic short answer grading[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009：567-575.

[46] 乔亚男，齐勇. 广义词汇共现模型研究[J]. 中国计算语言学研究前沿进

展, 2009(3)：13-15.

[47] 宗成庆, 吴华, 黄泰翼, 等. 限定领域汉语口语对话语料分析[C]//全国第五届计算语言联合学术会议论文集. 1999：115-122.

攻读硕士学位期间发表的学术论文及其他成果

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《网络购物环境下的问句答案匹配方法研究》，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：陈俊杰　　日期：2013 年 7 月 8 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：陈俊杰　　日期：2013 年 7 月 8 日

导师签名：　　日期：2013 年 7 月 8 日

致　谢

在论文即将完稿之际，谨向在这一过程中给予我指导和帮助的实验室老师，同学以及我的家人表示衷心的感谢。

回首研究生生涯，导师王晓龙教授和陈清财教授严谨细致的工作态度，平易近人的指导风格让我受益匪浅，为我在科研道路的行进树立了榜样，指明了方向。在此衷心感谢两位导师对我的关心和指导。

本课题的顺利完成，首先要感谢整个问答组的所有同学。科研的每一步进展，到最后的完成，每一步都是在他们的指导和帮助下完成的。他们给予了许多宝贵的建议和帮助。他们在学习科研中认真负责、一丝不苟的态度，是我以后学习工作中的榜样。感谢徐军博士，为本课题确立了研究方向和论文初期研究的无私帮助，对我细心的指导使我在学习和研究上得到巨大的积累。感谢候永帅博士，在课题的研究过程中对我实验方法及论文撰写提出了许多宝贵的意见和建议。感谢周小强博士，在课题初期的语料处理给予的细心指导。

同时，感谢同组的王一方同学，在我研究过程中遇到困难时一起分析问题解决问题。同感谢问答课题组的文博、刘岭岭师弟和李亚辉师妹，他们为课题组平台搭建、组内语料收集标注以及系统性能改进等各个工作上付出的努力。

本文的研究工作得到了实验室所有同学的大力帮助。感谢同期的同学刘增健、李昊迪、官山山、廖梦、豆荣刚、葛丽萍，他们在学习和生活中给予我很大的帮助。能够和你们身处同一个实验室，并度过人生最宝贵的硕士研究生生涯，是我的无比荣幸。在此向全体实验室成员致以最衷心的祝愿！

感谢一直呵护我成长的父母，是他们的爱帮助我坚信自己，鼓励和支持我度过每一次难关。

最后，在此向辛勤培育我成长的母校、哈尔滨工业大学研究生院和培养我的导师们表示深深的谢意！