

汉语情感问题类型分类研究

葛正荣 李婷玉 姚天昉

上海交通大学计算机科学与工程系 上海 200240

hh190@sjtu.edu.cn yao-tf@cs.sjtu.edu.cn

摘要: 随着网络搜索引擎技术的飞速发展, 对于问答系统的需求愈发迫切。而问答系统处理问题的第一步就需要分辨情感问题和非情感问题并对情感问题进行分类。本文首先分析了当前问答系统和问题分类领域的研究现状, 总结了一些存在的问题。然后针对情感问题从三个方面进行分类。在语义层面, 提取了三个关键词; 在语法层面, 通过规则的制定, 将其分成五种疑问句类型; 在领域层面, 通过搜索引擎的相关网页数量来进行判断。再综合上述三个方面开发出测试系统进行了分析。实验结果表明: 对于情感问题的分类, 从三个层面进行分析比较全面。

关键词: 情感问题, 问答系统, 问题分类, 自然语言处理

Research on Chinese Sentiment Question Category Classification

Zhengrong Ge Tingyu Li Tianfang Yao

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240

hh190@sjtu.edu.cn yao-tf@cs.sjtu.edu.cn

Abstract: With the development of search engine, the demand of question answering system becomes larger. The first step of QA system dealing with question is judging whether a question is an opinion question or a non-opinion question and classifying opinion questions. The paper firstly analyzes the current situation of QA system and Question Classification, and then summarizes some problems. We analyze the question in three layers. In the semantic layer, we extract three key words. In the syntactic layer, we classify the question into five typical question types by some rules. In the domain layer, we judge the domain by the number of related websites which is got through search engine. After then, based on three layers talked above, we develop a system for experiments and do further analysis. The result of experiments shows that for the classification of opinion question, analyzing it from three layers is very useful.

Key words: opinion question, QA system, question classification, Natural Language Processing

1 概述

网络技术的高速发展, 宣告了信息时代的到来。如何将自己需要的信息从网络上海量的数据中抽取出来, 已经成为一个热点研究方向。传统搜索引擎的方法往往是依靠关键字的搜索, 通过匹配来检索出用户所需要的信息和内容。但是, 关键字的匹配并不能很好地获取用户所需要的内容, 往往会夹杂着大量的无用信息, 需要用户人工筛选。而问答系统 (Question Answering) 的出现有望代替原有的信息检索技术, 使用户能够更方便地找到所要的信息。

传统的问答系统一般只针对比较简单的、基于事实的问句, 而随着大量评论性信息的出现, 传统的问答系统回答这些情感问题不再显得有效。对此, Wiebe 和 Cardie 等人首先实现了一个 MPQA (Multi-Perspective Question Answering) 系统, 它没有专门对问题进行分类,

而只是简单地采用人工的方法对情感问题和基于事实的问题进行手工标识, 构建了一个主观性文本语料库, 通过分析和抽取语料库中的情感信息片段, 回答情感问题。Someasundaran 和 Wilson^[1]等人在此基础上, 进一步对情感的类型体系进行细分。接着, Yu 和 Hatzivassibloglou^[2]分别在篇章层和句子层对情感分类做了研究。针对一个情感问题, 他们试图在文本中寻找与问题具有相同情感倾向的句子, 然后根据这些情感句子产生答案, 但是在情感问题分析方面, 仅仅只对情感问题的极性进行了分类。Kim 和 Hovy^[3]专门对情感问答系统中句子的情感持有者识别做了研究。Lun-Wei Ku 和 Hsin-His Chen^[4]等人基于 TREC-11 和 NTCIR 等语料实现了一个简单的情感问答系统。

情感问题分析是情感问答系统的第一个处理步骤, 它对于理解问题的关键信息、提高检索速度、缩小答案范围具有重要的意义。由于汉语的搜索技术原本就开展比较晚, 再加上中文的复杂性, 所以西方研究者所提出的一些方案与技术往往会“水土不服”。本文将重点研究非特定领域的情感问题分类, 从语法、语义、领域等不同层面分析并定义情感问题的类型。在语法层面, 采用汉语语法的疑问句式的分类标准; 在语义层面, 利用关键字等特征进行分类; 在领域层面, 可以采用本体等工具确定领域。在此基础上建立情感问题类型体系。接着筛选区分情感问题和非情感问题的敏感特征, 这些特征可以来源于不同的语言层面。在完成情感问题和非情感问题分类以后, 针对情感问题, 采用情感问题类型体系对它进行进一步分类, 确定问题的关键信息(如命名实体)和预期候选答案的关系。

2 情感问题类型分类

一般对基于示例的指导性学习进行问题分类, 都按照疑问短语来进行^[5], 其主要方法有以下几种: 统计方法、规则学习以及树方法、神经网络方法等。其中最常用的是统计方法, 其次是规则学习和树方法, 而神经网络方法比较少用。

在本次实验中, 我们考虑从语法、语义、领域等层面分析并定义情感问题的类型。在语法层面, 采用汉语语法的疑问句式的分类标准; 在语义层面, 利用关键字等特征进行分类; 在领域层面, 可以采用本体等工具确定领域。并在此基础上建立情感问题类型体系。针对情感问题, 采用情感问题类型体系对它进行进一步分类, 确定问题的关键信息和与预期候选答案的关系。

所以, 总的来说, 对于问题的分析, 最终将有三部分组成:

- a. 语义层面的一些关键词
- b. 语法层面的不同疑问句类型的分类
- c. 领域层面领域的确定

2.1 语义层面

对于一般的文本来讲, 往往可以通过一些关键词的搜索来进行语义层面的分类。但是问题分类有其特殊性, 由于篇幅较小, 一个问题不超过 20 个语素, 所以往往其语义不容易判断。而对于情感问题来说, 它问题的覆盖面更狭窄, 情感问题往往是围绕着一个情感表达事件来进行的提问, 因此只要将关于这个情感表达事件描述的关键词抽取出来, 那么, 整个情感表达事件的大致意思就可以了解。通过对大量情感问题的观察, 我们认为主要有以下几个关键词:

(1) 情感词 (sentiment word)

对于绝大多数情感问题来说, 极性的形容词或者动词往往是表达情感和看法用

的最多的语素。因为它带有一定的情感倾向，而这些倾向则往往用来表达人对于事物的态度。本次实验利用知网(HowNet)情感词辞典来判断情感词。值得一提的是，情感词具有两种词性，一为形容词，一为动词。由于这两种词在句中的的语法作用不同，所以往往要分开标记。

(2) 主题词(target)

在情感问题中，情感词是用来修饰主题词的，是为了对该主题表达一定的情感。在本次试验中，主要利用实验室已有的汽车主题库进行搜索^[6]。

(3) 持有者(holder)

在情感问题中，不同的人 and 团体对于相同的主题可能抱有不同的情感和看法，因此持有者也应该作为一个关键词。

由于持有者可以是任何人、组织或者团体，范围很大，所以利用辞典不可能囊括，因此还要加上语法的判断，比如动词前面很有可能就是情感的持有者，特别是建议性动词和情感动词。所以对于这两者前面的名词、代词、和一些专有名词，都认为是情感持有者并提取出来。要注意的是，很多句子中会缺少情感持有者。

例如：吉普车安全吗？

问题的情感持有者被省略了。通过分析，本次试验中把情感持有者省略的状况认为是对于“你”的提问。对于上例即看作“你认为吉普车安全吗？”

2.2 语法层面

本次实验中对于情感问题分类的语法层面研究向英语语法借鉴^[7]。将其分成两大类，即一般疑问句和特殊疑问句，鉴于一般疑问句的回答比较简单，所以不再细分。而特殊疑问句首先与上一节抽取到的三个关键词相对应。用英语来表示即为：情感词对应“how”，主题词对应“what”，持有者对应“who”。同样问原因“why”的特殊疑问句也可以放在情感问题的分类里面，但是对于“where”和“when”这两个问地点和时间的特殊疑问词，则将他们舍去，因为它们总是出现在非情感问题当中，不在我们所讨论的情感问题范畴之内。

情感问题分类图表如下所示：

表 2-1 情感问题分类表

| 提问内容 | 疑问词 | 示例 |
|------|--------|--------------|
| 情感词 | “how” | 马自达怎么样？ |
| 主题词 | “what” | 你喜欢什么车？ |
| 持有者 | “how” | 谁认为奥迪舒适？ |
| 原因 | “why” | 为什么女生喜欢迷你宝马？ |
| 判断 | | 奔驰车安全吗？ |

情感词类：这类问题是询问人或者团体对于某一事件或事物的看法，而这类问题回答的要点总是在于情感词，以下是几种基本的问句类型：

表 2-2 情感词类句式规则表

| 序号 | 例句 | 句式规则 |
|----|-----------------|------------------------|
| 1 | 大学生认为奇瑞 QQ 怎么样？ | 持有者+建议性动词+主题词+“how”疑问词 |
| 2 | 奇瑞 QQ 怎么样？ | 主题词+“how”疑问词。 |
| 3 | 奇瑞 QQ 大学生认为怎么样？ | 主题词+持有者+建议性动词+“how”疑问词 |

主题词类：这类问题是询问人或者团体什么事物或事件持有特定的看法，而这类问题回答的要点总是在于主题词，有以下几种基本的问句类型：

表 2-3 主题词类句式规则表

| 序号 | 例句 | 句式规则 |
|----|---------------|-------------------------|
| 1 | 欧洲人认为什么跑车豪华 | 持有者+建议性动词+“what”疑问词+情感词 |
| 2 | 欧洲人认为豪华的跑车是什么 | 持有者+建议性动词+情感词+“what”疑问词 |
| 3 | 什么跑车欧洲人认为豪华 | “what”疑问词+持有者+建议性动词+情感词 |
| 4 | 豪华的跑车欧洲人认为是什么 | 情感词+持有者+建议性动词+“what”疑问词 |
| 5 | 欧洲人喜欢什么跑车 | 持有者+情感动词+“what”疑问词 |
| 6 | 什么跑车欧洲人喜欢 | “what”疑问词+持有者+情感动词 |
| 7 | 什么跑车豪华 | “what”疑问词+情感词 |
| 8 | 豪华的跑车是什么 | 情感词+“what”疑问词 |

持有者类：这类问题是询问哪些人或团体对某些事物或事件持有某种看法，而这类问题的回答的要点总是在于持有者，有以下几种基本的问句类型：

表 2-4 持有者类句式规则表

| 序号 | 例句 | 句式规则 |
|----|--------------|------------------------|
| 1 | 谁认为奥迪舒适 | “who”疑问词+建议性动词+主题词+情感词 |
| 2 | 谁认为舒适的车是奥迪 | “who”疑问词+建议性动词+情感词+主题词 |
| 3 | 认为奥迪舒适的是谁 | 建议性动词+主题词+情感词+“who”疑问词 |
| 4 | 认为舒适的车是奥迪的是谁 | 建议性动词+情感词+主题词+“who”疑问词 |
| 5 | 奥迪谁认为舒适 | 主题词+“who”疑问词+建议性动词+情感词 |
| 6 | 舒适的车谁认为是奥迪 | 情感词+“who”疑问词+建议性动词+主题词 |
| 7 | 谁喜欢奥迪？ | “who”疑问词+情感动词+主题词 |
| 8 | 喜欢奥迪的是谁 | 情感动词+主题词+“who”疑问词 |
| 9 | 奥迪谁喜欢 | 主题词+“who”疑问词+情感动词 |

原因类：这类问题是询问情感事件表达的原因，而这类问题的回答要点总是在于理由，有以下几种基本的问句类型：

表 2-5 原因类句式规则表

| 序号 | 例句 | 句式规则 |
|----|----------------|----------------------------|
| 1 | 为什么女生觉得迷你宝马时尚 | “why”疑问词+持有者+建议性动词+主题词+情感词 |
| 2 | 女生为什么觉得迷你宝马时尚 | 持有者+“why”疑问词+建议性动词+主题词+情感词 |
| 3 | 女生觉得迷你宝马时尚是为什么 | 持有者+建议性动词+主题词+情感词+“why”疑问词 |
| 4 | 为什么女生喜欢迷你宝马 | “why”疑问词+持有者+情感动词+主题词 |
| 5 | 女生为什么喜欢迷你宝马 | 持有者+“why”疑问词+情感动词+主题词 |
| 6 | 女生喜欢迷你宝马是为什么 | 持有者+情感动词+主题词+“why”疑问词 |
| 7 | 1的情感词倒装 | “why”疑问词+持有者+建议性动词+情感词 |

| | | |
|----|------------|----------------------------|
| | | +主题词 |
| 8 | 2的情感词倒装 | 持有者+“why”疑问词+建议性动词+情感词+主题词 |
| 9 | 3的情感词倒装 | 持有者+建议性动词+情感词+主题词+“why”疑问词 |
| 10 | 为什么迷你宝马时尚 | “why”疑问词+主题词+情感词 |
| 11 | 迷你宝马时尚是为什么 | 主题词+情感词+“why”疑问词 |
| 12 | 10的情感词倒装 | “why”疑问词+情感词+主题词 |
| 13 | 11的情感词倒装 | 情感词+主题词+“why”疑问词 |

一般疑问句类：在情感问题中，一般疑问句常常询问某一情感表达事件的正确与否，而这类问题的答案一般只有两种：正确或错误。有以下几种基本的问句类型：

表 2-6 一般疑问句句式规则表

| 序号 | 例句 | 句式规则 |
|----|-------------|--------------------------|
| 1 | 你是不是认为奔驰车安全 | 持有者+A不(没)A+建议性动词+主题词+情感词 |
| 2 | 你认为奔驰车是不是安全 | 持有者+建议性动词+主题词+A不(没)A+情感词 |
| 3 | 你认为奔驰车安全不安全 | 持有者+建议性动词+主题词+A不(没)A |
| 4 | 你是不是喜欢奔驰车 | 持有者+A不(没)A+情感动词+主题词 |
| 5 | 你喜欢不喜欢奔驰车 | 持有者+A不(没)A+主题词 |
| 6 | 1的情感词倒装 | 持有者+A不(没)A+建议性动词+情感词+主题词 |
| 7 | 2的情感词倒装 | 持有者+建议性动词+情感词+A不(没)A+主题词 |
| 8 | 2的省略 | 主题词+A不(没)A+情感词 |
| 9 | 3的省略 | 主题词+A不(没)A |
| 10 | 奔驰车安全吗? | 陈述句+句尾疑问词 |

从上述五种疑问句类型中可以总结出几点情感问题句式规律：

- (1) 一般来说主题词和情感词的位置可以互换，且不影响句子的表达。
- (2) 一般来说句子带有持有者和建议性动词的，两者可以省略，不影响句子的表达。
- (3) 疑问词的不同和某种关键词的缺少往往决定了句子的类型。

2.3 领域层面

对情感问题的分析来看，决定这个问题是属于哪个领域最关键的是主题词。主题词往往决定了情感的对象，这个对象属于哪个领域则这个情感问题也属于该领域。在本次试验中，利用 google 来进行相关性的搜索。比如主题是“奔驰”，那么在 google 中输入“奔驰 汽车”，记录其相关网页数量，同时也输入“奔驰 食物”，记录其相关网页数量，“汽车”和“食物”为领域。当将主题词和不同的领域的搜索引擎相关网页数得到后，其中数值最大的领域便是这个主题的主题词。对于问题的领域，并不是说要将其限定在某个领域中，而是将其归纳到某个领域中方便之后的答案筛选等步骤，对于问题本身没有影响。

例如：对于主题“宝马”的领域判断，以下数据来源于 www.google.com

- 约有 19,100,000 项符合 宝马 汽车 的查询结果
- 约有 9,730,000 项符合 宝马 食物 的查询结果
- 约有 1,570,000 项符合 宝马 mp3 的查询结果

可见宝马和汽车的相关网页最多，所以判断宝马属于汽车领域。

3 系统介绍及实验分析

从下图来看，在语义、语法、领域三个方面，语义层次的分析最重要，它是一切的基础，关键词的标注是另外两步的基础。而语法层面和领域层面在拿到语义分析的结果后则可以分别独立执行，最后将三个部分的结果整合输出。

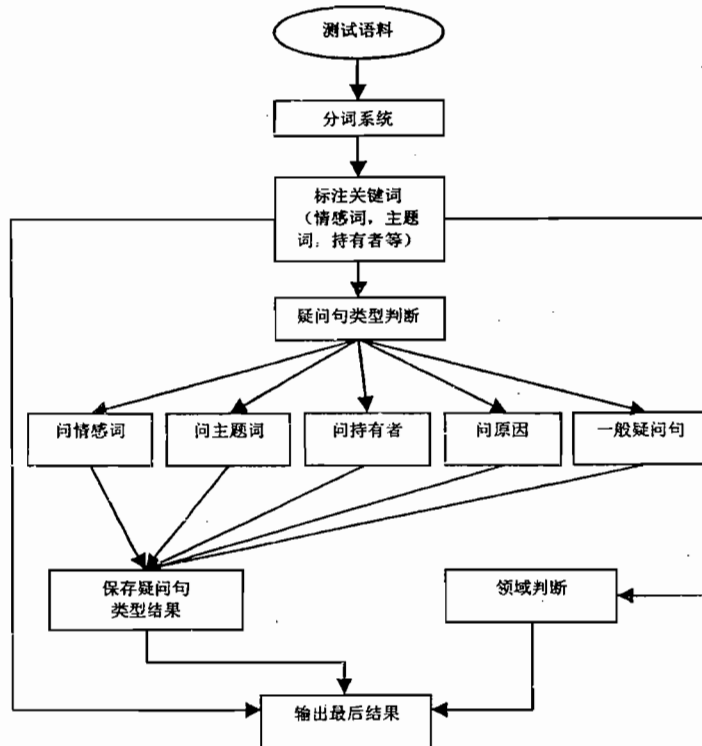


图 3-1 情感问题分类流程图

本次实验对于每种疑问句类型准备 20 句左右的测试语料，并按难度将其分成三等，对整个情感问题分类系统进行测试，并且手工判断其结果正确与否。最后的结果如下表所示：

表 3-1 情感问题分类系统实验结果

| 问题类型 | 问题难度 | 总句数 | 初次实验正确句数 | 修正后正确句数 |
|------|------|-----|----------|---------|
| 问持有者 | 简单 | 7 | 7 | 7 |
| 问持有者 | 中等 | 7 | 5 | 7 |
| 问持有者 | 困难 | 6 | 4 | 4 |
| 问主题词 | 简单 | 7 | 7 | 7 |
| 问主题词 | 中等 | 7 | 3 | 7 |
| 问主题词 | 困难 | 6 | 1 | 4 |
| 问情感词 | 简单 | 6 | 6 | 6 |
| 问情感词 | 中等 | 6 | 5 | 6 |
| 问情感词 | 困难 | 4 | 2 | 3 |
| 问原因 | 简单 | 7 | 7 | 7 |

| | | | | |
|-------|----|---|---|---|
| 问原因 | 中等 | 7 | 2 | 7 |
| 问原因 | 困难 | 6 | 2 | 5 |
| 一般疑问句 | 简单 | 6 | 5 | 5 |
| 一般疑问句 | 中等 | 7 | 3 | 6 |
| 一般疑问句 | 困难 | 6 | 2 | 4 |

从上表看来,本次实验的结果一开始不是十分理想,仔细观察一些错位的句子,发现主要有以下几个主要问题:

例 1: 哪些人觉得宝马车是最舒适的跑车?

这句问题,系统的类型判断正确,但是对于情感词的提取则不够精确,系统提取了“舒适”而不是“最舒适”,而“最”,“不”,“比较”这类副词对于后续研究中对于答案的搜索有着至关重要的作用,所以应该将情感形容词前的副词一并加入情感词输出。

例 2: 为什么人们觉得奥迪漂亮安全?

该句疑问句分类正确,但是对于情感词的提取不够全面,系统提取了“漂亮”,但是“安全”则被忽略了。在情感问题中,常常有这一类问题,其主题拥有多个情感词。应该将并列的情感词一起输出。

修改了以上一些问题之后,用新的系统来进行测试,结果有了很大的改善,对于简单和中等的情感问题处理得比较理想,但是,对于困难的情感问题,处理起来仍有些吃力,主要是由于汉语随意性较强,句型、词性等在不同的情况下都可以有不同的变化。

4 结束语

本文对情感问题分类研究,从语义,语法和领域三个方面对其进行分类。语义上挑选出三种关键词;在语法上,将问题分成 5 个类型,并制定相应规则,对其进行分类;在领域上,利用网络搜索引擎得到结果进行分类。实验的结果比较理想。经过试验,我们认为汉语情感问题分类最关键的部分还是在语法层面,如果能将每个词在句子中所起到的作用分析清楚并且在系统中很好的识别,分类会方便很多。

综上所述,可以看出,汉语情感问答系统作为一个新兴的搜索引擎发展方向,对其的研究才刚刚起步,不管是理论上还是系统开发,仍有很大的提升空间。

参考文献

- [1] SOMASUNDARAN S, WILSON T, WIEBE J, et al. QA with attitude: exploiting opinion type analysis for improving question answering in on-line discussions and the news. Proceedings of the International Conference on Weblogs and Social Media. Boulder, Colorado, USA, 2007.
- [2] YU H, HATZIVASSILOGLU V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan, 2003: 129-136.
- [3] Kim, S-M and Hovy, E. 2005. Identifying Opinion Holders for Question Answering in Opinion Texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*.
- [4] Lun-Wei Ku etc. Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems.
- [5] 郑实福等. 自动问答综述. 中文信息学报第 16 卷第 6 期
- [6] 姚天昉, 聂青阳, 李建超, 李林琳, 陈柯. 一个用于汉语汽车评论的意见挖掘系统[A]. 中文信息处理前沿进展-中国中文信息学会二十五周年学术会议论文集[C]. 北京: 清华大学出版社, 2006, 260-281.
- [7] 张刚等. 开放域中文问答系统的研究与实现. 哈尔滨工业大学信息检索研究室论文集 第一卷 2003.