汉语意见型主观性文本类型体系的研究*

刘全升　姚天昉　黄高辉　刘军　宋鸿彦

上海交通大学计算机科学与工程系 上海 200240

xubvlin@sjtu.edu.cn　yao-tf@cs.sjtu.edu.cn　fangao2620@163.com　steven.jun.liu@sjtu.edu.cn

songhy1985@hotmail.com

摘 要： 主观性文本是一种描述个人想法、情感和意见等的非约束性文本。它与主要描述以事实为主的客观性文本在内容和结构上有很大的不同。意见型文本是包含有意见元素（意见持有者、意见陈述范围、意见主题和意见情感）的一种主观性文本，它大量出现在网上的电子公告板、论坛和博客等媒介中，受到广泛的关注，并成为研究意见挖掘方法和技术的语料。在本文中我们介绍了主观性文本的定义及其与客观性文本的差异，同时着重讨论了意见型文本的定义、特点、类型体系及其在意见挖掘技术中的应用。由于对汉语主观性文本类型体系的研究才刚刚起步，有许多方面需要进行交流和商榷。我们在本文中讨论了汉语意见型文本的类型体系，意在抛砖引玉，希望得到同行的评价和指点，以利于把这项研究工作做得更好。

关键词： 主观性文本，意见型主观性文本，类型体系，意见挖掘

Study on the Category Architecture of Chinese Opinioned-subjective Text

Quansheng Liu　Tianfang Yao　Gaohui Huang　Jun Liu　Hongyan Song

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

xubvlin@sjtu.edu.cn　yao-tf@cs.sjtu.edu.cn　fangao2620@163.com　steven.jun.liu@sjtu.edu.cn

songhy1985@hotmail.com

Abstract: Subjective text is a non-restricted text that describes people's idea, emotion and opinion. It's very different from objective text that is used to state fact in content and structure. Opinioned text is a type of subjective texts, which contains opinion elements (holder, claim, topic, sentiment). It is very popular in BBS, forum and blog in the Internet, which attracts more and more attention and becomes the corpus for opinion mining. In this paper we introduce the difference between subjective text and objective text, and emphatically discuss the definition, character and category architecture of opinioned text and its application in opinion mining. The study on the category architecture of Chinese subjective text is just at his dayspring, there are many aspects of the need for exchange and discussion. Our study accounts for a modest spur to induce you to come forward with valuable comment. We hope we can obtain the evaluation and guidance from you so that we can do well in this investigation.

Keywords: subjective text, opinioned-subjective text, category architecture, opinion mining

1　概述

意见挖掘技术自 20 世纪 90 年代以来，随着因特网的兴起，一直受到来自学术界，产业界等的多方关注，意见挖掘技术处理的对象是语言上不受限制的真实文本，其主要来源为因特网上

* 本文承国家自然科学基金项目（项目号 60773087）和上海交通大学中德语言技术联合实验室的资助。

的大量非约束性文本，即主观性文本。相比于客观陈述，即描述事实的客观性文本来说，主观性文本在内容中多多少少总含有说话人的自我成分[1][2][3]，在结构上随意，主要表现为语言上的不受限性。当前对主观性文本的研究重点是一种包含有意见元素的文本-意见型主观性文本（以下简称意见型文本）。在意见型文本中抽取信息相对比较容易。随着研究的深入，利用意见型文本进行情报收集，商业信息提取等方面有越来越多的应用。但作为意见挖掘研究对象的意见型文本的研究相对较少，这影响了对意见挖掘方法和技术的研究。本文的工作就是在这样的背景下开展的。我们针对意见型文本中语句的特点介绍了以意见元素为基础的意见型文本的类型体系，为研究意见挖掘方法和技术所使用的语料进行分析与选择打下基础。

目前，对主观性文本以及意见型文本的语言现象以及相关方面的应用，国内外已有相关的研究。中国社会科学院语言研究所沈家煊等人讨论了自然语言的发展，语言的主观性与主观化趋势等问题[1]；美国南加洲大学 Kim、Hovy 等人定义了适合计算机处理的意见型文本的意见四元素以及讨论了主观性文本的随意性和处理上的困难性[4]。在利用主观性文本进行意见挖掘的应用方面，微软美国研究院 Gamon 等人开发了可以自动挖掘网上用户所上载的有关汽车评价信息的 Pulse 系统[5]，美国伊利诺斯大学 Liu 等人开发了可以处理网上在线顾客产品评价的原型系统[6]，上海交通大学姚天昉等人开发了用于汉语汽车评论的意见挖掘系统 Surveyer[7]。

2 汉语主观性文本

2.1 主观性文本的定义与特征

“主观性”是指语言的这样一种特性，说话人在说出一段话的同时表明自己的立场、态度和感情，从而在话语中留下自我的印记。一般来说，说话者多多少少在说话的时候总带有主观含义。主观性文本主要用于描述个人、群体、组织等的想法、观点与情感等非事实内容，因此具有强调自我成分与非约束性文本等特点。在内容上，主观性文本与客观性文本有着显著的不同。客观性文本主要描述客观的人、物、事等，主要以陈述句为主。主观性文本主要表达说话者对某人、物和事等的看法，因此主观性文本强调自我表达的内容。在内容表达上，主观性文本极具个性化，自由无约束。因此主观性文本中用词和句型无限制，文本中常常出现非规范词语和非规范句法结构等。因此，这是导致其在意见挖掘中出现困难的主要原因。人们早就已经注意到语言的“主观性”。但各种语言具有的“主观性”是不同的。有的语言表现“主观性”的形式很明显，例如日语，说日语时几乎不可避免地要用明确的语言形式来表达说话人对所说内容和对听话人的态度或感情。像英语这样的拉丁语系语言“主观性”的表现方式比较隐晦，但仍然大量存在。同上述语言相比，汉语的主观性主要表现在词语的用法上，而不是表现在词语的意思上。例如汉语动词重叠的主观性表达，虚词的使用等等。动词重叠有加重主观性的意思[8]。

例 1：“她小心地碰了碰她的手指，打落了烟蒂。”“她小心地碰了手指，打落了烟蒂。”

在上例的两句句子中，前者的主观性明显比后者强。从对比中我们可以发现在汉语中用动词重叠比不用动词重叠更能显示出主观性来。

2.2 主观性文本的表现形式

自然语言是人类用来交际的语言，在使用自然语言时，或多或少总能留下使用者对人、物和

事的主观看法，因此含有主观性内容的文本广泛存在于我们的生活中。在现实中纯客观的文本比较少，一般在客观性文本中总含有一些主观性的句子。

例 2：“1949 年，中华人民共和国成立。”为描述客观事实的客观性句子，它与常出现于互联网的主观性句子有明显的不同。客观性文本一般描述内容规范，客观严谨。

例 3：“强烈谴责 CNN 主持人恶毒辱华！”“做人不能太 CNN！！！”从这两句语句可以看出主观性文本通常具有口语式的特点，描述自由，常常也不规范。它往往含有非规范的词语或者语法结构。

3 意见型文本

对意见型文本的研究是目前意见挖掘的主要研究方向。意见型文本是包含有意见元素（意见持有者、意见陈述范围、意见主题和意见情感）的一种主观性文本。在意见型文本中，通常包含有表达意见（即具有褒贬成分）的语句。

3.1 意见型主观性文本的定义

根据 Kim 和 Hovy 对意见的定义，意见由四个元素组成：即主题(Topic)，持有者(Holder)，陈述(Claim)和情感(Sentiment)。这四个元素之间存在着内在的联系，即意见持有者针对某主题发表了具有情感的意见陈述。Kim 和 Hovy 对意见的定义既适合语言学阐述，又适合计算机处理，因为意见挖掘的过程就是要在自然语言文本中自动地确定这些元素以及它们之间的关系。意见挖掘的主要子任务如主题抽取，意见持有者识别，陈述选择，情感分析等都是针对组成意见四元素的。因此这种定义可以明确意见挖掘的目标，适合计算机处理意见型文本。包含意见型语句的文本就是意见型文本，意见型语句主要是指包含意见元素的语句。但有时候一个意见型句子中，不一定存在所有的意见元素，如意见持有者就常常缺失。

3.2 意见型与非意见型文本

意见型文本是主观性文本中具有代表性的一类自然语言文本，但并非所有意见型文本都具有主观性。

例 4：“张三听李四说李四很不喜欢鹅。”

在例 4 的句子中“李四很不喜欢鹅”是一个意见型语句，因为其中包含意见，具有意见四元素（主题：“鹅“，意见持有者：“李四”，情感“很不喜欢”，陈述“李四很不喜欢鹅”），但对整句话而言，是一个包含意见的客观陈述句，因此是客观性语句，它描述了客观事实。所以不一定所有的意见型文本都是主观性的。相反，主观性语句一般都包含一些说话者的自我意见。

例 5：“我认为攻打伊拉克将会使美国陷入困境。”“我喜欢这车。”

在上述两句句子中，前者隐含了对“攻打伊拉克”的否定情感，后者包含了对“车”的显式的肯定情感。因此在处理时，前者提取不出明显的对主题的情感，而后者可以容易地抽取出意见四元素。

3.3 意见型文本类型体系

目前对意见型文本分类的研究较少，主要是因为目前对意见型文本的研究主要是基于意见

挖掘的应用驱动的。这样就造成了从语言学的角度来深入分析意见型文本的类型体系也相对较少。本节试图从语言学的角度来对意见型文本的类型体系进行初步的归纳。

3.3.1 意见型文本的来源

意见型文本主要用来表达作者（说话者）对某人、物或事的情感，媒体是人们表达情感、发表观点比较集中的地方。网络、报刊、书信、短信、影评等均比较容易出现具有褒贬意见成分的意见语句。需要指出的是，在众多媒体中，互联网的迅速发展是使意见型文本作为意见挖掘的主要对象以及意见挖掘方法和技术研究活跃的主要原因。互联网的发展使任何组织和个人均可以利用网络在法律的约束下自由发表自己的言论。随着论坛、个人博客、聊天工具的兴起与普及，互联网上的“意见”呈几何级数增长，普通人发表意见和表达愿望随着互联网的发展变得越来越容易。因此，在各种意见型文本的来源中，网络上的意见型文本是最为丰富及多种多样的。意见挖掘方法和技术的研究所涉及的意见型文本主要也是来自于互联网上的文本信息。

3.3.2 意见型文本的表现形式

互联网上意见型文本具有多种表现形式，可以以评论形式出现，也可以以聊天记录的形式出现。根据互联网上意见型文本表现形式的不同，可将其分为小说、散文、诗歌、报刊评论、读者来信、新闻报道、采访记事、科学论文、正规书评/影评/诗评、日记、作文、邮件、信函、博客、论坛、BBS、读者书评/影评、个人主页、聊天记录、贴吧等等。下面将对互联网上具有代表性的表现形式进行举例分析。

例 6：“她记得在夏威夷接受日裔移民官审查时，那人脸上谨慎严肃的表情。他是个拥有权力决定张爱玲未来的人。他眼睛梭巡着张爱玲，一边问些套话，一边对她进行主观的考量。她只能保持着低调诚恳的态度，即使说到被留在身后的亲人时心头轻轻有些抽搐，也必须抑制住从眼神里流露出的丝毫情感。”

例 6 中句子摘选自网上小说《她从海上来》，从中可以看出网上小说句子结构较为规范，同时意见型语句“含量”较少。

例 7：“祖国的好儿女！我为你们骄傲为你们自豪！！！13 亿中国人支持你们！！！”“楼主我顶啊！！！”“我们会在国内努力工作、建设祖国、振兴中华！！ 抵制一切洋垃圾，做一个有尊严的中国人！！”例 7 中三句句子摘选自中华网论坛，从中可以看出，论坛中意见型文本来源丰富，文本中意见型语句比例较高，情感明显、强烈，但句子中容易出现非规范词以及非规范的句子结构，增加了计算机处理的难度。论坛中谈论的话题一般来自于单一领域，适合目前的意见挖掘技术处理。

例 8：“反暴力、反污蔑！明天（19 日），一场海外华人爱国大游行将在欧洲大陆和北美拉开帷幕。”上例中句子来自于新快报报道，从中可以看出作为意见型文本表现形式的新闻报道的一些特点，新闻报道文本内容比较规范，文本来源丰富，领域多样，但文本中含有意见型语句没有论坛多，情感一般较为适中。新闻报道适合作为舆情分析的文本，用于进行舆情分析。

例 9：“我喜欢诺基亚N系列的手机，因为它的功能中有一个记事本功能是我最喜欢的。在日常生活中，我喜欢用我的N73 手机记录下我生活中的所见所闻及人生感悟。”该句子选自新浪博客中的一个个人主页，个人主页是个人自由抒发意见，表达看法的地方。因此从个人主页中往往可以看到许多意见型文本，但其包含意见型语句的比例不一，会随着主人的情感，习惯而变化。同时个人主页书写自由，可能出现规范抒情的散文式语句，也可能出现论坛中常有的强烈情

感倾向的意见型语句。个人主页常常也是多领域的，因此目前比较难以作为意见挖掘的处理对象。

在上述例子中我们对小说，论坛，新闻报道，个人主页等主观型文本的表现形式作了分析，通过对互联网中各种表现形式的分析，可以找出适合作为意见挖掘处理对象的文本。无论选用何种类型的文本，使用时都需要经过适当的预处理才适合作为一般数据挖掘处理的对象。

3.3.3 意见型文本的结构

相比较于客观性文本来说，主观性文本一般不规范。但意见型主观性文本包含种类繁多，来源亦不一，因此意见型主观性的各类文本，其结构的规范性是不一样的。在此主要讨论互联网上意见型文本的规范性。互联网上意见型文本具有多种意见表现形式。在上述表现形式下意见型文本的规范性有所不同，按计算机处理时的难度，可将上述形式下的意见型文本分为三类，即规范程度比较高的文本，程度一般的文本（即可能规范也可能不规范），不规范的文本。如下表所示：

文本规范程度	文本表现形式
规范	小说、散文、诗歌、报刊评论、读者来信、新闻记事、采访报道、科学论文、正规书评/影评、诗评
比较规范	日记、作文、邮件、信函
不规范	博客、论坛、BBS、读者书评/影评、个人主页、聊天记录、贴吧

表 1： 按文本规范程度划分的意见型文本类型体系

需要指出的是，互联网上以各种表现形式出现的意见型文本，均可能出现规范程度不一的现象，上述的分类只是大致按一般情况下使用计算机处理时的处理难度进行分类。

随着互联网的发展，网络非规范语言[9]的发展也越来越迅速，网络非规范语言是指用户在使用网络时派生或创造出来的表达自己意思的一种独特的交往语言。如例 10：“细八细这样地？”“斑竹是猪！”这两句句子中均出现了网络非规范语言。前者中“细八细”是规范语言中“是不是”的谐称。后者中“斑竹”表示“版主”，表示特定论坛版块的管理者。意见型文本中常常出现网络非规范语句，如上例中第二句就表达了意见，是意见型语句。目前一般的意见挖掘过程中预处理阶段经常需要包含对网络非规范语言的处理，以便能够正确理解语句中非规范词汇的词义。在讨论意见型文本的类型体系时，不可避免地要考虑网络非规范语言对意见型文本分类的影响，按网络非规范语言在意见型文本中的“含量”不同，可将意见型文本划分为以下两类：

（1） 含网络非规范语言较多的意见型文本：论坛，BBS，聊天记录，贴吧；

（2） 含网络非规范语言较少的意见型文本：正规报道，评论，散文，诗歌，信函。

据《第 19 次中国互联网络发展状况统计报告》[10]显示，截至 2006 年底，我国网民人数达到了 1.37 亿，占中国人口总数的 10.5%，中国互联网将迎来快速增长期。如此庞大的网民群体大范围使用网络非规范语言将对我们的规范母语形成一定的冲击，同时对自然语言处理技术提出了新的挑战。如何处理迅速流行起来的网络非规范语言也是意见挖掘过程中必须要面对的。

3.3.4 意见型文本的内容

正如人们说话的方式多种多样一样，利用主观性文本进行意见表达的种类也是多种多样的，因此出现了不同类型的意见，如有主题意见或无主题意见，具有多种情感的意见等，现举例说明。

例 11：“小毛病太多，送我也不要！”“我非常喜欢中华轿车。”“外型漂亮，但内部配置太差了。”例 11 共有三句句子，第一句表达了强烈的否定情感，但没有主题，即我们不知道

说话者的情感是针对什么的。因为说话者的随意性，这种句子在意见型文本中经常出现。第二句中“中华轿车”是说话者的主题，“非常喜欢”是说话者对主题的显式情感，而且主题和情感都只有一个。在意见型语句中也会常常出现多主题或多情感的意见，如第三句例句。因此通过以上的例句可以说明，意见型语句中主题与情感的出现是比较随意的，正是由于说话者的随意性，使意见型语句中意见四元素的组合多种多样。

从内容上讨论意见型文本，我们还可以针对意见型文本中意见针对的领域来探讨，意见型文本自由的表达形式可以针对多种领域。目前互联网上的各种论坛大都针对特定领域，如单个领域的针对某类产品的意见，评论等，现举例说明：

例 12：“惠普笔记本外观漂亮，性能不错，价格公道。”“我既反对藏独势力干扰奥运，也反对抵制家乐福，更反对 CNN 的不实报道！”例 12 句中前者描述的对笔记本电脑的评论，后者是对各种事件的评论。前者涉及的是单领域的主题，而后者明显涉及了多个领域。单领域意见型文本多见于产品评论，特定事件评论中，涉及多领域的文本常常出现于综合性论坛中。利用意见挖掘技术处理意见型文本时，处理单领域文本与处理多领域文本的难易程度是不一样的。目前意见挖掘中使用的本体学习，基于规则的匹配，机器学习等算法大都是领域相关的。在一个领域中具有较好效果的算法在另一个领域不一定能取得同样的效果。目前意见挖掘的多种应用如产品评论信息抽取，舆情分析等都基于特定的领域。

此外，意见的表现形式不同，互联网上出现的意见型文本丰富程度也不同，意见型文本的丰富程度与意见挖掘的效果密切相关。从意见型文本的内容上按两种方法分类结果如下表所示：

分类形式	文本表现形式	
内容领域	单领域	产品论坛、读者书评、科学论文
	多领域	综合论坛、BBS、小说、散文、诗歌、日记、作文、报刊评论、读者来信、采访报道、博客
文本可获取数量	多	论坛、BBS、读者书评/影评、博客、个人主页/空间、聊天记录、贴吧
	一般	日记、作文、报刊评论、读者来信、采访报道、新闻记事、科学论文、正规书评/影评、邮件
	少	小说、散文、诗歌、诗评、信函

表 2： 从内容上讨论主观性文本类型体系

3.4 适合作为意见挖掘处理对象的文本

意见挖掘技术需要使用意见型文本作为处理对象。一般来说需要较大规模的文本进行处理，目前意见型文本中语言现象常常不规范，意见型语句和非意见型语句混杂、部分意见元素、隐式主题、主题与情感元素存在多对多的关系。这些都不利于计算机的处理。意见挖掘处理对象的选择可以对互联网上意见型文本的各种表现形式进行分析，目前独立选取的语料可以推荐各种论坛出现的意见型文本。适合作为语料的文本应该是来自于互联网上的文本，即电子版的文本，而且在文本可获取的数量上应该具有一定的规模。意见型文本的规范性直接影响到利用计算机处理的难度，因此文本的句法结构和语义的难度应该控制在目前意见挖掘技术能够处理的范围内。同时文本所涉及的领域应该是多方面的，例如使用两到三种独立领域类型的文本以及一种开放领域类型的文本，且独立领域类型的文本的处理性能是可比较的，这样可以对处理结果进行比较，评价

挖掘的效果。另外，适合作为意见挖掘处理对象的文本内应该含有较高比例的意见型语句。这些条件都将使所选择的意见型文本适合作为意见挖掘的处理对象，能够使挖掘取得更好的效果。

4 结束语

在本文中我们首先介绍了从事这项工作的动机和目标，即旨在更有效地研究汉语主观性文本，特别是意见型文本。我们的目标是通过研究汉语意见型文本的类型体系，为意见挖掘技术所处理语料的选择作好准备。其次我们从语言学的角度介绍了主观性文本的定义，特点以及汉语主观性文本的一些特性。最后我们详细地介绍了汉语意见型文本的类型体系以及互联网上意见型文本的各种表现形式，通过所给出的综合实例说明了意见型文本各种形式的特点。同时对适合作为意见挖掘语料的条件作了分析，讨论了适合作为语料的意见型文本的表现形式，我们介绍了相关的研究工作并对它们作了一定的评价。以上主要是对意见型文本的类型体系研究作了一个介绍。这方面的研究可以说还刚刚开始，总体上还不够深入，但值得引起我们的重视。汉语中主观性文本的特点，意见型文本的类型，互联网上新出现的网络非规范语言的处理，何种类型的意见型文本适合于作为意见挖掘使用的语料。对这些问题我们都应该在国内外已有研究成果的基础上作进一步的探索。我们在这篇论文中探讨了汉语意见型文本的类型体系，希望得到国内同行的评价和指点，以利于把这项研究工作做得更好。

参 考 文 献

[1] 沈家煊，语言的主观性与主观化[J]. 外语教学与研究，2001(4)

[2] Lyons, J.Semantics. Cambridge: Cambridge University Press. 2vols

[3] Finegan.dward. Subjectivity and subjectivisation: an introduction. In D.stein& S.wright 1995.

[4] S.-M. Kim and E.Hovy. Determining the Sentiment of Opinions [A]. In: Proceedings of COLING-04, the Conference on Computational Linguistics (COLING-2004) [C]. Geneva, Switzerland: 2004, 1367-1373.

[5] M. Gamon and A. Aue. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms [A]. In: Proceedings of the ACL- 2005 Workshop on Feature Engineering for Machine Learning in NLP [C]. Michigan, USA: 2005, 57-64.

[6] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews [A]. In: Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI-2004) [C]. San Jose, USA: 2004.

[7] 姚天昉，聂青阳，李建超，李林琳，陈柯，付宇. 一个用于汉语汽车评论的意见挖掘系统[A]. 见曹右琦，孙茂松主编，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.

[8] 潘国英. 论汉语动词重叠的主观性表达[J]. 修辞学习，2008(1)

[9] Y. Xia, K-F. Wong, and W. Li. A Phonetic-Based Approach to Chinese Chat Text Normalization [A]. In: Proceedings of the 21st International Conference on Computational Linguistics and 44 Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006) [C]. Sydney, Australia: 2006, 993-1000.

[10] 第 19 次中国互联网络发展状况统计报告. 中国互联网络信息中心 （http://www.cnnic.net.cn/）