

非监督的汉语感情语料库的构建及分析

陈瑛 李逸薇 黄居仁

香港理工大学 中文及双语学系

E-mail: {chenying3176,sophiaym,churen.huang}@gmail.com

摘要: 目前感情计算是个全新的研究问题, 该问题首先面临的是缺乏相关的语料库。本文提出一种非监督的方法用以创建大规模的文本语料库。该语料库包括一个感情句子语料库和一个中性(无感情)句子语料库。鉴于非监督的方法的自由及方便性, 我们可以快速创建大规模的语料库用来帮助感情计算的研究, 从而避免了繁重的人工标注。同时, 本文在借鉴西方的感情分类体系的基础上, 提出中文的感情分类体系, 并对中文的感情词进行归类以及对感情的数据分布做了一些定量分析。

关键词: 感情, 感情语料库, 语料库的建立

Construction of Chinese Emotion Corpus with an Unsupervised Approach

Chen Ying, Sophia Y. M. Lee, Huang Chu-Ren

Department of Chinese & Bilingual Studies, the Hong Kong Polytechnic University, Hong Kong

E-mail: {chenying3176,sophiaym,churen.huang}@gmail.com

Abstract: emotion computing is a new topic for NLP, and the first problem for this task is the lack of annotated data. Here, we provide an unsupervised approach to create a corpus, which includes an emotion corpus (the sentences which contain a kind of emotion expressions, such as “happiness”, “anger”, and so on) and a neutral corpus (the sentences which contain no emotion expression). One big advantage of our method is that we can create a very large emotion corpus very quickly so that to avoid time-cost human-annotation. Because of few previous work on Chinese emotion classification, here, we borrow some theories, which deals with English emotion classification, to classify Chinese emotion words. Finally, we gave the data analysis for the corpus we create.

Keywords: emotion, emotion corpus, corpus creation

1 前言

近几年, 随着 sentiment 的深入研究, 感情(emotion)分析也同时受到越来越多的关注。相比 sentiment 分析, 感情分析可以克服它的很多局限性。首先, 感情是人对外界事物的一种本能反应, 所以感情分析可以克服 sentiment 的领域局限性。再次, 与 sentiment 相比, 感情分析更深入更细致(不局限于“好的”和“不好的”), 因而能给计算语言学的应用系统提供更多的信息。

感情研究是个非常复杂的问题。迄今为止, 无论是从语言学的角度还是从计算语言学的角度, 汉语的感情研究都比较少。文献[1]对汉语的情感¹做了一些语言学理论上的分析, 他们列出了情感词及其相应的情感类。但是我们也发现文献[1]对其情感分类并没有提供相应的理论支持, 因而其情感词的归类是没有系统性的。文献[2]列出了汉语的一个感情树, 但也没有给出相应的理论根据。在本文, 我们对汉语感情分类体系做进一步的研究和完善。根据 Turner 的感情分类体系[3], 我们对汉语的感情词进行系统地归类, 这无论是对感情的分析, 如感情的合成和分解, 还是对感情计算, 都是非常重要的。

¹ 在计算语言学中, 情感指的是 sentiment。在汉语中, 情感是一个比较泛指的概念, 它包括感情(emotion), 态度(attitude), 品性(personal trait)等。

感情计算通常包括两方面：感情识别（区分无感情和有感情的）和感情分类（对感情作进一步的分类）。感情分类需要大型的感情语料库（对有感情的文本进行类别标注）。遗憾的是，迄今为止，很少有学者去标注大型的感情语料库。值得一提的是文献[2]和[4]。文献[4]利用博客中的 emoticon 信息自动地建立起一个与感情互现的词库，但单纯的词库会散失很多与感情相关的语境信息，不利于感情计算。文献[2]提供了一个人工标注的感情语料（包含 15449 无感情句和 24039 感情句），但是，这个数量对感情计算是远远不够的。标注大型的语料库需要大量的时间和精力，其中涉及语料库的选择、收集与整理，标注体系的制订，以及语料的标注。而且，感情标注往往因人而异，具有很强的争议性，这会导致语料库的质量很难得到保证。另一方面，感情识别通常还需要无感情句子语料库，即中性句子语料库。本文从做感情计算的实际出发，用非监督的方法创建了一个大型的汉语感情语料库和中性句子语料库用以满足感情计算的需要。由于非监督的方法的自由应用性，我们采用了两个不同的语料库去标注感情句，并给出一些感情分布的定量分析。

2 汉语感情语料库

语料库往往是针对某些特定的应用而创建的，应用方式的不同往往会影响语料库的整个创建方式。因此，我们首先分析一下感情语料库的一些可能的应用。感情语料库的建立是为了感情计算。如果一个句子含有感情关键词，从直观上来说，它的感情类别是很容易识别的（通过收集感情关键词并对其进行归类）。但实际上，由于词语的歧义性，这种简单的识别实际上并不能取得理想的效果。例如，“紧张”既可以用来形容人的感情（例如，他很紧张这件事），有可以用来描述气氛等（例如，气氛一下变得紧张起来）。但是，相比含有感情关键词的句子，从那些不含有感情关键词的文本里识别出含有感情的句子辄是更难的任务。由于感情问题的复杂性，本文采用 Wierbicka 的 Natural Semantic Metalanguage 理论[5]来作为感情计算的理论基础：感情往往是由某个外界刺激而引起的，例如，“环境污染”往往伴随着“担忧”，“攻击”往往伴随着“愤怒”。这样，感情计算可转化为对引起感情的刺激事件的识别。因此，感情语料库的创建就转化为对引起感情的刺激事件的收集。但是，由于感情具有“自明性”[6] - “主体对自身情感状态的感受却没有与他人商讨或者征求意见的必要”，单纯的刺激事件语料库并不能完全解决感情计算问题。例如，“兴建球场供大家运动打球”在不同的情况下可以引起不同感情，如“惊讶”，“高兴”等。同时，我们也注意到在一个含有感情词的句子中，往往会给出该感情的刺激事件及其语境，这就指出了一个简单的收集感情刺激事件的方法，即创建感情语料库的方法。接下来我们先讨论一下感情标注体系，包括感情分类体系和感情词的选择，然后再给出非监督的感情标注方法。由于感情计算的需要，我们也建立了一个高质量的中性句子语料库。

2.1 汉语感情分类

感情的古典定义是“感情是一种心理状态，是当身体因受到观察到或想到的事件影响而有所反应的自觉”（Emotion is the felt awareness of bodily reactions to something perceived or thought）[7]。在西方的感情研究中，很多学者分别从生物学、进化学、哲学、心理学等不同角度，对感情(emotion)给出了不同的定义（具体见[8]）。

虽然不同文化对感情的表达和理解会有所不同，但是大多数西方学者都认同有些感情是最普遍的（基本感情）。很多西方学者对基本感情类别的选定存在不同的看法，但“happiness, anger, sadness, fear”（乐怒哀惧）基本上是公认的四种基本感情[9]。中国古代学者对感情分析比较少，文献[1]对汉语的基本感情做了一些总结：

“在文人学者的著作中，先秦时代，中国的学者主要将“情”分为六种或七种类型……著名的文艺理论家多统一到“七情六欲”的说法上……唐代韩愈在《原性》中说：“其所

以为情者七：曰喜、曰怒、曰哀、曰惧、曰爱、曰恶、曰欲。”……中国传统上涉及的基本感情类型应可归纳为七类，即：“好（爱，敬）、恶，喜（乐），怒，哀，惧，欲”等”除了基本感情，其他的感情都可归为复合感情。

对于感情语料的标注，除了感情的定义之外，感情的分类也是非常重要的。在东方很少有学者研究感情词的整体分类，而在西方相关的研究却比较多，因此，本文借鉴了西方的感情分类体系。但是我们也注意到东西方语言中的感情词并非是一一对应的映射，为此我们针对汉语做了一些相应的改进。在感情分类方面，Plutchik模型[8]是当代西方最有影响力的感情分类体系，Turner模型[3]，[10]，[11]是在Plutchik模型的基础上的进一步完善。本文以Turner的感情分类体系作为整个汉语感情语料库标注的理论基础。Turner的感情分类模型认为：

1) 可把四种基本上公认的感情，“happiness, anger, sadness, fear”（乐怒哀惧），作为基本感情。同时，他也承认恶（disgust）和惊（surprise）也是可以作为基本感情的[10]，[11]。例如，“惊”在[10]中就被定义为一种基本感情。考虑到“惊”在汉语是一个比较常见的字，并有很强的成词能力，所以在我们的汉语分类体系中，我们将“乐怒哀惧惊”作为基本感情。同时，像Plutchik模型一样，Turner还把每种基本感情按照强烈程度分成强、中、弱。

2) 复合感情可由几个基本感情合成，并可分解成一级感情（有两个基本感情合成，一个是主要的，另一个是次要的），二级感情（有三个基本感情合成）等。文献[11]给出了由“乐怒哀惧”合成的一些常见的一级和二级感情词。同时，我们参考文献[10]加入了含“惊”的一级感情词。

给定了感情分类体系，对于每一个感情词，我们需要选择其隶属的感情类别。由于篇幅的限制，本文只对文献[1]中的374个基于心理感受的情感词进行归类。首先，我们去掉12个单音节词（因为单音节词比较容易引起歧义）。对剩余的每个情感词（有362个情感词），我们让两个标注者进行判断，只有在两者都同意的情况下，这个词才被判定为感情词。最后，我们试着把这些感情词归类到Turner的感情分类体系中。表1列出了属于5种基本感情的感情词，表2给出了属于一级和二级感情的复合感情词。在表2中，我们把一些类似的感情词归为一类，例如，“妒嫉，妒忌，嫉妒，嫉恨，眼红，忌恨，忌妒”归为“嫉”的一类。同时，对每个复合感情，其所含的基本感情的顺序也代表了这些基本感情在合成这个复合感情时所占的比重。如“嫉”由“惧”（主要）+“怒”（次要）合成。表1和表2仅仅列出了文献[1]中的189个基于心理感受的情感词，被剔除的173个词或者不是感情词，或者有歧义（具体语义要看具体的语境）。其实，对感情词的界定在学术上也一直有争议。例如，文献[11]认为“好（爱，敬）”不是感情词（emotion）。在我们实际的感情分类过程中，我们也发现有些词是处于模糊或灰色地带，如“闲适”，“放松”。同时，我们也加入了两个关于“欲”的感情词：“想要”和“想”。当这两个动词后面紧跟着一个动词的时候，它们往往表达的是“欲”的感情。

基本感情	感情词
Happiness/乐	强：痛快，振奋，亢奋，兴奋 中：欣慰，高兴，愉悦，欣喜，欢欣，开心，欢快，快慰，舒畅，快乐，快活，畅快，舒坦，欢娱，如意，喜悦，顺心，爽心，晓畅，松快，欢愉，欢喜 弱：怡和，自在
Fear/惧	强：惶恐，恐慌，恐惧 中：畏怯，心虚，心慌，害怕，发慌，发憷，苦恼，困惑 弱：羞怯，羞惭，害羞，害臊，迟疑，为难
Anger/怒	强：愤怒，忿恨，激愤，愤懑，愤慨，忿怒，悲愤，暴怒，蔑视，瞧不起，轻蔑，鄙夷，鄙薄，鄙视，自卑，痛恨，怨恨，憎恶，愤恨，厌烦，腻烦，恼恨

	中: 生气, 窝火, 厌倦, 讨厌, 厌恶, 反感, 冲动 弱: 烦躁, 烦乱, 烦心, 烦人, 烦杂
Sadness/哀	强: 悲恸, 悲痛, 哀伤, 悲哀, 哀痛, 沉痛, 痛心, 悲凉, 悲凄, 悲伤, 哀戚, 绝望 中: 感伤, 伤心, 伤感, 心酸, 憋气, 郁悒, 郁闷, 憋闷, 阴郁, 怅惘, 消沉, 颓丧, 颓唐, 烦闷 弱: 灰心, 丧气
Surprise/惊	强: 骇怪, 骇异, 震惊 中: 诧异, 吃惊, 愕然, 惊讶, 惊奇, 惊诧, 惊愕, 奇怪 弱:

表 1 属于基本感情的感情词

	组合	感情词
一级感情	乐+惧	傲: 骄横, 骄慢, 骄矜, 骄傲, 自信, 自豪, 炫耀, 得意 宽心: 宽心 欲望: 想要, 想
	乐+怒	解恨: 解恨
	乐+哀	感动: 感动
	惧+怒	嫉: 妒嫉, 妒忌, 嫉妒, 嫉恨, 眼红, 忌恨, 忌妒
	惧+哀	急: 焦虑, 焦渴, 焦急, 焦炙, 心浮, 心焦, 揪心, 心急, 着急 愁: 愁苦, 忧愁, 发愁, 担忧, 担心, 犯愁, 忧虑
	惧+惊	'惊恐: '惊骇, '惊惶, '惊惧, '惊恐, 吓人, 慌张, '惊慌, 惶惑, 慌乱
	怒+惧	疑: 怀疑, 疑心, 疑惑
	哀+惧	悲愁: 悲愁, 哀愁, 愁闷, 惆怅 后悔: 惭愧, 抱歉, 对不起, 羞愧, 对不住, 背悔, 懊恼, 懊悔, 悔恨, 懊丧
	哀+怒	委屈: 委屈, 冤枉, 抱委屈, 哀怨 不满: 不满, 不快, 不悦
	哀+惊	失望: 失望, 心寒
惊+乐	'惊喜: '惊喜	
二级感情	哀+惧+怒	疚: 负疚

表 2 属于复合感情的感情词

2.2 汉语语料库

对一个感情语料库而言, 所标注的语料库的文体类型对感情的分布是有很大影响的。一般来说偏文学的语料往往含有较多的感情信息。本文分别采用 Chinese Gigaword Corpus²和 Sinica Corpus³来标注感情。其中 Sinica Corpus (有 1 千万个词) 是一个平衡语料库, 它含有的文章涉及 15 种文体, 其中文学色彩比较丰富的文体, 如小寓言故事和散文, 占到 18.6%, 所以 Sinica Corpus 是一个比较适合标注感情的语料库。Chinese Gigaword Corpus 是个非常庞大的纯新闻语料库, 大约含有 12.9 亿个词。其语料分别是大陆 (新华社 XIN_CMN), 台湾 (中央社 CNA_CMN),

² <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T03>

³ <http://www.sinica.edu.tw/ftms-bin/kiwil/mkiwi.sh>

新加坡（联合早报 ZBN_CMN）三地收集而来的，具有汉语的地区使用性差异。另外，这两个语料库都做了分词和词性标注[12][13]，其中 Sinica Corpus 的分词和词性标注是经过人工纠正的。

2.3 非监督的汉语感情语料库创建

首先，我们给出感情语料库的非监督创建方式。在给定一个语料库和一个感情词词库及其感情分类体系的情况下，我们的感情语料库创建包括四个步骤：

1) 对每一个感情词，我们用正则表达式去提取含有该感情词的句子。同时，我们还要求这个感情词必须是动词，这可通过检查该感情词的词性标注来达到。而且，对跟“欲”有关的感情词（“想要”和“想”），我们进行了特别处理，即要求紧跟其后的词也必须是动词。

2) 我们再用正则表达式去发现这个感情词前面的词是否是否定词。如果是，则把这个句子剔除出去；

3) 每个句子的感情类别可以根据该句子含有的感情词给定（参照表 1 和 2）；

4) 在做感情计算时，句子中的该感情词会被删除掉，这样感情语料库只能为感情计算提供相应的语境。

为了保证感情语料库的质量，我们不但把有歧义的感情词去掉（在收集感情词词库时），而且把感情的否定句也删除掉（在步骤 2）。这主要是因为感情的否定句歧义性比较大。例如，

(1) 卡斯楚从来没有喜欢过俄国人，他认为他们驻军及经援他的国家是为了苏联自己的利益。

(2) 每次煮鸡汤都放一大堆中药，那我很不喜欢吃。

(3) 他指出，虽然华夏表示向中广借的钱都有还，可是每次借新债还旧债，钱越借越多，中广员工怎么能不担心。

句子(1)的“没有喜欢”其实并没有表达特别多的感情，所以这个句子是中性的。句子(2)中的“不喜欢”表达了“喜欢”的反面感情，所以这个句子是含有感情的。句子(3)中的“怎么能不担心”虽然是否定，但它表达的是一个肯定的意思，所以这个句子是含有感情的。总体来说，感情的否定可以表达三种意思：1) 只是表达了没有该感情的存在，即中性的；2) 表达了与该感情对立的一种感情；3) 进一步肯定了该感情的存在。如何对与感情变迁有关的句式（[14]），（如否定句，假设句式等）进行处理是我们下一步的工作。

感情的识别往往需要无感情句子语料库，即中性句子语料库。文献[15]提出对感情计算来说中性句子语料库的创建是非常重要的，而且中性句子的收集并非易事。比如，在日语的网络语料中大约有 80% 的句子是有感情的[15]。为了了解中性句子在汉语书面表达中的分布，我们采用一个比较直观的方法来自动地提取中性句子。给定一个句子，如果该句子及其前后句都不含有任何一个感情词（见表 1 和 2），那么该句子被归为中性。最后，感情语料库和中性句子语料库共同组成整个以感情计算为目的的语料库。

3 汉语感情语料库的分析

如果不考虑基本感情的强度差别，我们的分类体系总共包含 22 个感情类。表 3 给出感情语料在 Sinica Corpus 和 Chinese Gigaword Corpus 的分布。首先，从表 3 我们可以发现在 Sinica Corpus 中，最多的三个感情都是“乐”、“欲”、“怒”，而在 Chinese Gigaword Corpus 中，最多的三个感情都是“乐”、“不满”、“欲”。由于我们对每个感情类只列出了一些比较常用的感情词，所以表 3 只是给出了包含这些感情的句子的大致分布。同时，我们也注意到在不同的语料库，感情的分布是不同的。在 Sinica Corpus 中，我们自动提取的感情句子占到 3.3%，

而在Chinese Gigaword Corpus中,则只占到1.2%。这又证实了偏文学的语料库含有较多的感情表达。同时,我们从Sinica Corpus和Chinese Gigaword Corpus分别抽1214个和1238个句子进行人为标注。表4给出了它们的准确率。这里,我们采用了两种不同的评价方式。

- 1) 完全匹配评价: 只有自动标注和人为标注完全一致的情况下,该自动标注才是正确的;
- 2) 部分匹配评价: 由于次级感情是由几种基本感情合成,如果标注的感情(无论是自动标注和人为标注)是一个次级感情,它就被分解成其含有的几种基本感情。如果自动标注的任何一个基本感情在人为标注中出现,我们就认为自动标注是正确的。比如,自动标注的是“乐”,而人为标注的是“傲”(乐+惧),我们认为自动标注是对的。

从表4,我们看到自动提取的感情句的准确率大概在70%左右。这对一个有22个感情类的标注来说是可以接受的,同时,这也说明了感情是一个比较复杂的问题。当我们进一步分析自动标注的错误时,我们发现大概90%的错误句子其实际上的正确标注是“中性”。这一方面可以解释为什么上述两种评价方式得出的准确率差别不大,另一方面也说明了感情识别(区分有无感情)的重要性。

对中性句子的自动提取,我们从Sinica Corpus和Chinese Gigaword Corpus分别抽取995个和1006个句子进行人工标注,发现自动提取的准确率是都在98%以上。这说明我们对中性句子的自动提取方式是非常有效的。同时,我们也计算了一下中性句子的分布,在Sinica Corpus中,我们自动提取的中性句子占到48.9%,而在Chinese Gigaword Corpus中,中性句子占到65.5%。这同文献[15]的发现(最多有19.7%的日语网络语料是中性的)很不同。这个分布上的对比,可能来自两个原因。一是网络的非正式文本中,带感情的句子比例比较高。另一个可能是,文献[15]对感情的定义过于宽松。由信息理论以及带感情的文本通常带有重要信息这两个出发点看;文本中如果80%的句子带感情,则中性句子反而带有突出信息,这样的信息结构并不合理。所以我们的发现更为合理。

感情类	Sinica (句子数)	Gigaword (句子数)	感情类	Sinica (句子数)	Gigaword (句子数)
不满	581	48849	悲愁	54	194
傲	710	11446	愁	1482	37412
哀	1245	12733	欲	4391	47283
乐	4536	52251	感动	545	3955
失望	355	13132	怒	1708	7889
委屈	178	2179	疑	888	17269
嫉	114	604	疚	1	1
安心	9	212	解恨	0	4
惧	1299	19727	惊喜	112	1344
后悔	609	4623	惊恐	354	4167
急	428	4219	惊	1345	11247

表3 感情的分布

	完全匹配评价	部分匹配评价
Sinica Corpus	68.4	69.4
Chinese Gigaword Corpus	72.5	72.9

表4 自动标注的感情句的准确率

4 结论

感情计算是一个复杂而又全新的课题, 本文主要探讨了以感情计算为目的的语料库的建立。首先我们采用西方的比较成熟的感情分类研究来进一步完善汉语的感情分类体系。然后, 以感情理论为基础, 我们给出了一个非监督的语料提取方法来自动创建一个大规模语料库。该语料库包含一个具有一定准确率的感情语料库和一个高质量的中性句子语料库。

同时, 我们也发现了许多问题需要进一步研究和探讨。对于感情分类体系, 如何划分基本感情和复合感情, 复合感情的合成和分解, 及感情词的选择和归类都是值得进一步研究的问题。不同的应用系统往往会关注不同的感情类别, 设计一个灵活的感情分类体系对于感情计算来说是非常重要的。对于感情语料库, 如何对感情词和感情句进行进一步的筛选以保证感情语料库的准确率和覆盖率, 在不同的语料库中中性句子和感情句子的分布等等, 都需要进一步的探讨, 而且这些都会对感情计算产生直接的影响。同时, 我们也注意到以上这些问题会涉及到不同的学科, 且存在很大的争议性, 所以在下一步的工作中, 我们会着重于从感情计算的角度来考虑和解决这些问题。

参 考 文 献

- [1] 许小颖, 陶建华, 汉语情感系统中情感划分的研究, 会议第一届中国情感计算及智能交互学术会议第一届中国情感计算及智能交互学术会议论文集, 2003.
- [2] 徐琳宏, 林鸿飞, 赵晶, 情感语料库的构建和分析, 中文信息学报, 2008.
- [3] Jonathan H. Turner. On the origins of human emotions: A sociological inquiry into the evolution of human affect. Stanford, CA: Stanford University Press. 2000.
- [4] Changhua Yang, Kevin Hsin-Yih Lin and Hsin-Hsi Chen, Building Emotion Lexicon from Weblog Corpora, Association for Computational Linguistics. 2007.
- [5] Anna Wierzbicka. Semantics: Primes and Universals. Oxford: Oxford University Press. 1996
- [6] 赵春利, 情感形容词与名词同现的原则, 中国语文, 2007.
- [7] Walter B. Cannon. The James-Lange theory of emotion: A critical examination and an alternative theory. American Journal of Psychology, 39:10-124. 1927.
- [8] Robert Plutchik. Emotions: A psychoevolutionary synthesis. New York: Harper & Row. 1980
- [9] Theodore D. Kempter. How many emotions are there? Wedding the social and autonomic components. American Journal of Sociology, 93, 263-289. 1987.
- [10] Jonathan H. Turner. The Evolution of Emotions in Humans: A Darwinian-Durkheimian Analysis. Journal for the theory of social behaviour 26:1-34. 1996.
- [11] Jonathan H. Turner. Human Emotions: A sociological theory. New York: Routledge, 2007.
- [12] Wei-Yun Ma and Keh-Jiann Chen. Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. In Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, 168-171. 2003.
- [13] Wei-yun Ma and Chu-Ren Huang. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy. 24-28 May. 2006.
- [14] Livia Polanyi and Annie Zaenen. Contextual Valence Shifters. In James G. Shanahan, Yan Qu, Janyce Wiebe (Eds.), Computing Attitude and Affect in Text: Theory and Applications, pp. 1-10. 2004.
- [15] Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion Classification Using Massive Examples Extracted from the Web. COLING. 2008.