SenticNet 4：语义资源
基于概念基元的情感分析

erik cambria, soujanya poria,
Soujanya bjpai
南洋理工大学新加坡南洋大道50
号
{坎布里亚，sporia
rbajpai} @ntu。

bjo-rn schuller
伦敦帝国理工学院，皇后
大道180号，赫胥黎大厦，
伦敦SW7 2AZ，英国

bjoern.schuller@imperial.ac.uk

摘要

传统人工智能系统与人类智能之间的一个重要区别在于人类利用从一生的学习和经验中获取的常识知识来做出明智决策的能力。这使人类能够轻松适应由于缺乏特定情境规则和泛化能力而导致AI灾难性失败的新情况。常识知识还提供背景信息，使人们能够在通常假设这种知识的社交场合中成功地运作。由于常识包含人类认为理所当然的信息，因此收集它是一项极其困难的任务。以前版本的SenticNet专注于收集这种情绪分析知识，但是由于无法概括而严重受限。SenticNet 4通过利用通过层次聚类和降维来自动生成的概念原语来克服这些限制。

1 介绍

由于市场营销和财务预测带来的巨大好处，这引起了许多激动人心的公开挑战（Pang），因此获得公众意见的机会引起了科学界和商界的兴趣。和Lee，2008; Liu，2012）。然而，从自然语言挖掘意见和情感是一项极其困难的任务，因为它需要深入理解语言的大多数明确和隐含，规则和不规则的句法和语义规则。现有的情绪分析方法主要依赖于明确表达意见的文本部分，例如极性术语，影响词及其共现频率。然而，意见和情感通常通过潜在的语义隐含地传达，这使得纯粹的句法方法无效。

SenticNet（Cambria等，2014）根据语义和语义来捕获这些潜在信息，即通常与现实世界对象，行为，事件和人相关的外延和内涵信息。SenticNet不再盲目地使用关键词和单词共现计数，而是依赖于与常识概念相关的隐含意义。优于纯粹的句法技巧，SenticNet可以通过分析不明确传达情感的多字表达来检测巧妙表达的情绪，而是与这样做的概念相关。SenticNet的主要限制是它无法概括概念的实例，例如吃意大利面或啜食面条：除非有完全匹配，否则SenticNet 3会引发一个未发现的错误。

然而，在SenticNet 4中，动词和名词概念都与原语相关联，因此，例如，吃意大利面或啜食面条等概念被概括为INGEST FOOD。通过这种方式，知识库可以捕获大多数概念变形：像吃，啜食，蒙克等动词概念都由它们的概念原语INGEST表示，而面食，面条，牛排等名词概念则用它们的本体父母FOOD代替。

第26届COLING 2016会议论文集，第26届计算语言学国际会议：技术论文，第2666-2677页，日本大阪，2016年12月11日至17日。

这种概括背后的想法是，有一组有限的心理基元用于影响承载的概念和一组有限的心理组合原则来控制它们的相互作用。通过层次聚类和降维的集合应用，在SenticNet中自动发现概念原语。

本文的其余部分安排如下：第2节介绍情感分析领域的相关工作;第3节提出了概念原语的概述;第4节和第5节分别详细描述了名词概念和动词概念的概括性;第6节提出了两个不同的最先进数据集的实验结果;最后，第7节提供了结束语。

2 相关工作

情感分析系统可以大致分为基于知识的系统或基于统计的系统（Cambria，2016）。虽然知识库的使用最初在识别文本中的情感极性方面更受欢迎，但最近情绪分析研究人员越来越多地使用基于统计的方法，特别关注监督统计方法。庞等人。（Pang et al。，2002）通过比较电影评论数据集中不同机器学习算法的性能并获得82％的极性检测准确度，开创了这一趋势。Socher等人最近的一种方法。（Socher et al。，2013）使用递归神经张量网络在同一数据集上获得了85％的准确度。Yu和Hatzivassiloglou（Yu和Hatzivassiloglou，2003）使用词的语义方向来识别句子级别的极性。梅尔维尔等人。（Melville et al。，2009）开发了一个框架，利用词类关联信息进行依赖于域的情感分析。

最近的研究利用微博文本或特定于Twitter的功能，如表情符号，主题标签，URL，@符号，大写和伸长，以增强推文的情绪分析。唐等人。（Tang et al。，2014a）使用卷积神经网络（CNN）来获取在推文中经常使用的单词的嵌入。Dos Santos等。（dos Santos和Gatti，2014）也专注于深度CNN，用于短文中的情感检测。最近的方法还侧重于开发基于情感语料库的单词嵌入（Tang et al。，2014b）。这些单词向量包括比常规单词向量更多的情感线索，并且对于诸如情绪识别（Poria等，2016b）和方面提取（Poria等，2016a）之类的任务产生更好的结果。

然而，统计方法通常在语义上较弱（Cambria和White，2014）。这意味着，除了明显的影响关键字之外，统计模型中的其他词汇或共现元素单独具有很小的预测价值。因此，统计文本分类器在给定足够大的文本输入时仅以可接受的准确度工作。因此，虽然这些方法可能能够在页面或段落级别上有效地对用户的文本进行分类，但它们在较小的文本单元（例如句子）上不能很好地工作。相反，概念层面的情感分析侧重于通过使用网络本体或语义网络对文本进行语义分析，这允许聚合与自然语言观点相关的概念和情感信息（Cambria和Hussain，2015; Gezici et al。，2013; Arau'jo et al。，2014; Bravo-Marquez et al。，2014; Recupero et al。，2014）。

通过依赖大型语义知识库，这些方法远离盲目使用关键词和单词共现计数，而是依赖于与自然语言概念相关的隐含特征。与纯粹的句法技巧不同，基于概念的方法也能够检测出以微妙方式表达的情感;例如，通过分析没有明确传达任何情感的概念，但这些概念与其他概念隐含地相关联。概念包模型可以表示与自然语言相关的语义，比词袋更好。事实上，在后者中，像丑陋或悲伤的微笑这样的概念会被分成两个单独的词，破坏输入句子的语义和咒语。

3 概念原语

试图对事物，事件和人物进行分类，寻找共同的模式和形式是人类固有的。关联两个实体的最直观方式之一是通过它们的相似性。根据格式塔理论（Smith，1988），相似性是指导人类对世界的感知的六个原则之一。

相似性是使一个人或另一个人像另一个人一样的品质，“相似”意味着具有共同的特征。根据颜色，形状，大小和纹理等因素，可以通过多种方式将对象视为相似的。如果我们远离纯粹的视觉刺激，我们可以应用相同的原则来定义基于共享语义特征的概念之间的相似性。先前版本的SenticNet利用这一原理来聚类具有相似情感特性的自然语言概念。但是，查找类似概念的组并不能确保完全覆盖多字表达式的所有可能的语义变化。

在这项工作中，我们利用这些相似性推导出可以更好地概括SenticNet常识知识的概念原语。这种概括的灵感来自于概念原语的不同理论，包括Roger Schank的概念依赖理论（Schank，1972），Ray Jackendoff关于解释性语义表征的工作（Jackendoff，1976），以及Anna Wierzbicka关于素数和普遍性的书（Wierzbicka，1996），还有知识表示的理论研究（Minsky，1975；Rumelhart和Ortony，1977）。所有这些理论都声称需要一种分解方法来探索概念化。以同样的方式，物理学家通过将物质分解为逐渐变小的部分来理解物质，通过将意义分解为更小的部分来进行概念化的科学研究。显然，这种分解不可能永远持续下去：在某些时候我们必须找到无法进一步分解的语义原子。这是概念结构的层次;通过原始概念元素编码基本理解和常识的心理表征，其中构建了意义。

在SenticNet中，这种“分解”转化为多字表达的泛化，这些表达传达一组特定的情绪，因此具有特定的极性。这种泛化过程背后的动机是，有无数种方法可以用自然语言表达相同的概念，并且拥有所有可能的概念变形的综合列表几乎是不可能的。虽然可以通过词形还原来解决诸如共轭和变异之类的词汇变形，但是诸如同义词或语义相关概念的使用之类的语义变形需要通过类比推理来解决。如果在文本中遇到诸如获取知识和获取专有技术的多字表达式，则SenticNet 3无法处理它们，因为知识库中没有此类概念的条目。但是，SenticNet 3确实包含一个多字表达式，它与这两个概念在语义上高度相关，即获取知识。通过在原始层面上工作，Sen-ticNet 4能够弥合这一语义鸿沟，获取知识，获取专有技术，以及获取知识都由相同的概念原语表示：获取信息。

通过自动推断SenticNet概念的概念原语，我们的目标是广泛扩展常识知识库的覆盖范围，更好地执行情感分析任务，如极性检测和文本情感识别。如接下来的两节所示，这是通过层次聚类推广名词概念以及通过降维来发现动词概念的概念原语来完成的。

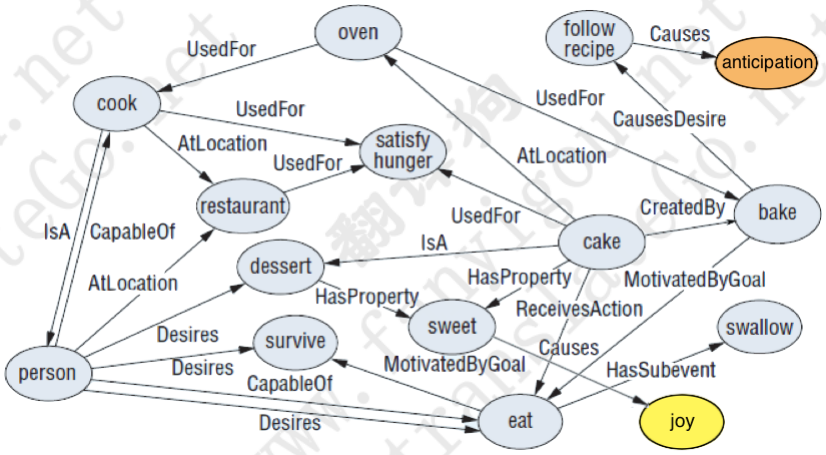


图1：AffectNet图的草图，显示了概念蛋糕的语义网络的一部分。

4 名词概念概括

在SenticNet中推广多词表达式的第一步是建立其名词概念（或对象概念）的层次分类，以便可以将猫，狗或宠物等名词识别为动物。通过在常识知识的语义网络上应用分层聚类来实现这种分类。值得注意的是，每个泛化都会继承情感信息和实例概念的极性。例如，在猫和狗的情况下，原始实际上是动物+因为猫和狗与积极情绪相关联。相反，对于与恐惧（例如，白鲨）或厌恶（例如蟑螂）等负面情绪相关的动物，相应的原语是ANIMAL-。

4.1 情感网

AffectNet （Cambria和Hussain， 2015）是一个建立在Con-ceptNet上的情感常识知识库（Speer 和 Havasi， 2012）， Open Mind 语料库的图形表示， 以及 WordNet-Affect （Strapparava和Valitutti， 2004）， 一种语言学影响的词汇表示的资源（图1）。资源表示为语义网络，其中节点是常识知识的多字表达，并且这些之间的链接是互连它们的关系。随着新版本的ConceptNet不断发布，新的情感常识知识通过游戏众包，AffectNet编码的知识也在不断扩展。首先将AffectNet转换为矩阵，将每个断言分为两部分：一个概念和一个特征，其中一个特征就是断言，第一个或第二个概念未指定，例如“轮子是’的一部分’或’是’一种液体’。

结果矩阵中的条目是正数或负数，根据断言的可靠性分配，其大小与置信度得分呈对数增加。因为AffectNet图是基于格式<concept-relationship-concept>的三元组构成的，所以整个知识库可以被视为一个大矩阵，一些语句的每个已知概念都是一行和每个已知的语义特征（关系+概念）是一个专栏。这种表示具有几个优点，包括执行累积类比的可能性（Tversky，1977），通过首先选择输入概念的一组最近邻居（就相似性而言），然后通过将该集合的已知属性投影到未知属性上来执行。概念（表1）。

4.2 集团平均凝聚聚类

动词+名词概念中的直接对象，例如买蛋糕或吃汉堡，表现出语义连贯性，因为它们倾向于生成具有相关语义的词汇项和短语。与同一动词相关的大多数单词倾向于共享一些语义特征。我们基于常识的方法类似于人类在寻找类似物品时所采取的方法 – 我们看看这些物品的含义有什么共同之处。在AffectNet中，概念彼此相互定义，有向边指示概念之间的语义依赖性。

概念	语义特征 (*关系+概念*)							
	..	*原因* 喜悦	*IsA* 事件	*用于* 家政	*位于* 派对场地	*部分* 庆典	*目标单位激励* 整理房间	..
.		.	.	.	.	.	.	
婚礼	..	0.94	0.86	0	0.79	0.88	0	..
扫帚	..	0	0	0.83	0	0	0.87	..
买蛋糕	..	**?**	**0.78**	**0**	**0.80**	**0.91**	**0**	..
生日	..	0.97	0.85	0	0.99	0.98	0	..
扫地	..	0	0	0.79	0	0	0.91	..
.		.	.	.	.	.	.	

表1：累积类比允许通过比较类似的概念来推断新的知识，例如，购买蛋糕导致快乐，因为婚礼和生日（类似的）这样做。

在语义网络中为任何特定概念c定义特征的传统方式是考虑通过c的出站边缘可到达的概念集。所提出的算法利用层次聚类从这些特征生成概念原语，其表示每个概念的核心语义。基于各种聚类算法的实验，例如k均值和期望最大化聚类，我们确定群平均聚集聚类（GAAC）提供最高的准确性。GAAC将数据划分为包含子集和兄弟集群的树（Berkhin，2006）。它生成树状图，指定不同级别的嵌套数据分组（Jain和Dubes，1988）。在聚类期间，概念表示为从AffectNet提取的常识特征的向量。接近矩阵构造为行和特征作为列的概念。如果要素是概念的出站链接，则矩阵中的相应条目为1，在其他情况下为0。余弦距离用作距离度量。集合算法本质上是自下而上的。GAAC尤其包括以下步骤：

1. 计算邻近矩阵。每个数据项都是一个初始集群。

2. 从邻近矩阵，通过合并形成一对簇。更新邻近矩阵以反映合并。

3. 重复，直到合并所有群集。

根据所需簇的数量，在高度修剪得到的树形图。群集之间的群组平均值由群组之间的平均相似距离给出。两个聚类之间的距离和相似性度量由下面的等式给出：

$$X_{sum} = \frac{\rightarrow^{-} C_n . C \rightarrow 米}{厘米\omega_{ii} \quad c_n \ \omega_i \ \omega_i, \ c_n / = c_m} \tag{1}$$

$$sim\ (z_i, \ z_j = \frac{1}{(n_i + N_j)(N_i + N_j - 1)} X_{sum} \tag{2}$$

其中 $^{-}$c是长度c概念的向量，向量条目是布尔值（如果存在特征，则为1）否则为0，并且N_i，N_j 分别是ω_i 和ω_j中的特征数（表示簇）。层次聚类算法的主要缺点是其运行复杂度（Berkhin，2006），其平均为 θ （$N^2 \log N$）。我们选择使用平均链路聚类，因为我们的聚类是基于连接的。概念邻近矩阵由AffectNet的特征组成，当两个概念共享多个特征时，“良好”连接被认为发生。聚类后，确定簇的数量并相应地修剪树形图。

随后将每个群集拆分为正子群集和负群集。群集实例根据其在SenticNet 3中的极性分配给正或负子群。例如，在应用GAAC后，眼镜蛇和猫最终处于同一群集（ANIMAL）中，但是因为它们在SenticNet中具有相反的极性在图3中，它们随后被分配到不同的子群集（分别为ANIMAL和ANIMAL +）。未发现特定分类的名词概念被分组在三个最常见的名词原语之一，即：SOMETHING，SOMEONE和SOMEWHERE（也分为正子集和负子集）。表2提供了24个名词概念的极性驱动的基于特征的聚类结果的示例。

事情-	东西+ 动物+	有些人- 专业的-	有人+ 专业+	某处- 性质-	某处+动物 - 自然+
蟑螂	马	掘墓人	医生	干草原	绿洲大鼠
	猫	验尸官	科学家	沙漠	沙滩
眼镜蛇	小狗	刽子手	老师	野生森林	自然公园白蚁

表2：极性驱动的概念原语推断的基于特征的聚类的示例。

5 动词概念泛化

概括SenticNet概念的第二步是为动词概念（或动作概念）定义概念原语，以便例如可以将获取，获得或收集等动词识别为GET。通过在AffectNet的向量空间表示上应用维数减少技术来实现这种分类。与名词概念一样，动词概念也与极性相关联，但在这种情况下，极性与这些动作概念所代表的相反含义（或结果）更相关，如增加与减少。这允许关于动词+名词组合的推理是按照代数乘法，其中负乘以正（或反之亦然）导致负，例如，减少增益（或增加损失），乘以两个正数产生正数，例如，增加PLEASURE，负数乘以负数会产生积极的结果，例如减少痛苦。

5.1 情感空间

人类思维通过不断压缩重要关系来构建可理解的意义（Fauconnier和Turner，2003）。压缩原则旨在将扩散和扩展的概念结构转换为更集中的版本，以便它们可以变得更适合人类理解。为了模仿这样一个过程，我们使用简单但功能强大的元算法来构建神经元学习（Lee et al。，2011）。这些元算法应该快速，可扩展，有效，几乎没有特定的假设，并且在生物学上是合理的。优化在过去几百万年的演化过程中形成的所有$\approx 10^{15}$连接是不太可能的。然而，客观地说，大自然可能只会优化全球连通性（主要是白质），但其他细节则随机化（Balduzzi，2013）。

在这项工作中，我们使用随机投影（Bingham和Mannila，2001）对AffectNet的矩阵表示进行压缩，以压缩与常识概念相关的语义特征，从而更好地对这些概念进行类比推理。随机投影是一种数据遗忘的方法，通过使用高斯N（0,1）矩阵将原始高维数据集映射到低得多的子空间，同时保持高概率的成对距离。这个理论上坚实且经验验证的陈述遵循Johnson和Lindenstrauss的引理（Balduzzi，2013），

这表明，对于所有点对x，y∈X，概率很高：

$$\sqrt{\frac{m}{d}}\,\|x-y\|_2(1-\varepsilon) \leqslant \|\Phi x-\Phi y\|_2 \leqslant \sqrt{\frac{m}{d}}\,\|x-y\|_2(1+\varepsilon) \tag{3}$$

其中X是欧几里得空间中的一组向量，d是该欧几里德空间的原始维数，m是我们希望将数据点减少到的空间的维数，ε是一个容差参数，用于测量允许的最大失真程度。度量空间，Φ是随机矩阵。在（Sarlos，2006）中引入了用于使矩阵乘法更快的结构化随机投影。当特征的数量远远大于训练样本（dn）的数量时，二次抽样随机哈达玛变换（SRHT）是首选，因为它的行为与高斯随机矩阵非常相似，但加速了从（nd）到（n log d）的过程。时间（Lu et al。，2013）。以下（Tropp，2011; Lu等人，2013），对于$d = 2^p$（其中p是任何正整数），SRHT可以定义为：

$$\Phi = \sqrt{\frac{d}{m}}\,RHD \tag{4}$$

其中•m是我们想要随机从d个特征中取样的数字；

• R是随机m×d矩阵。R行是来自R^d标准基础的m个均匀样本（无替换）；

• H∈$R^{d\times d}$是H_d2个正常1的Walsh-Hadamard matrix，它是递归定义的：

$$H_d = \begin{bmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{bmatrix} \quad \text{用} H_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix};$$

• D是ad×d对角矩阵，对角元素是iid Rademacher随机变量。

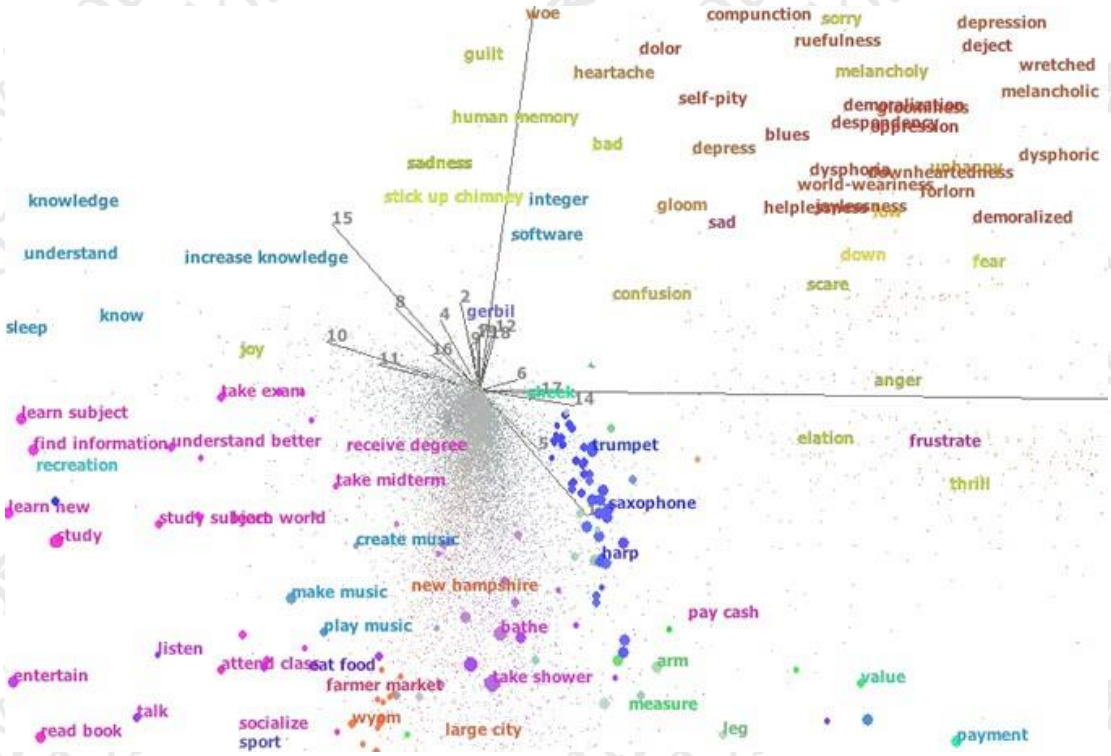


图2：在AffectiveSpace中，常识概念倾向于积极和消极的情绪。

我们随后的分析仅依赖于矢量对之间的距离和角度（即欧氏几何信息），并且足以将投影空间设置为数据大小的对数（Ailon和Chazelle，2010），因此，申请SRHT。结果是AffectiveSpace（Cambria等，2015），一个向量空间模型，其中常识概念和情感由m坐标的向量表示（图2）。

通过利用随机投影的信息共享属性，具有相同语义和情感效价的概念可能具有相似的特征 - 即，传达相同意义和情感的概念往往在AffectiveSpace中彼此接近。相似性不依赖于概念在向量空间中的绝对位置，而是取决于它们与原点的角度。例如，生日聚会，庆祝和购买蛋糕等概念在矢量空间中非常靠近，而失去信仰，沮丧和流泪的概念则在完全不同的方向上找到（与中心几乎相反）空间）。

5.2 半监督动词传播

AffectiveSpace允许对多字表达进行类比推理，以便检测购买杂货和购物等概念在语义上相似。然而，为了概括动词概念，我们需要根据动作区分这种推理，以便像购买杂货这样的概念与动词购买相关的概念相关联，例如购买牛奶或购买蔬菜。

为此，我们利用现有最大的英语动词词典VerbNet（Schuler，2005）和Sentic LDA（Poria等，2016c），这是一种在线性判别中计算单词分布的常识的分类框架。分析（LDA）算法。特别是，我们使用Sentic LDA的半监督版本，以便以这样一种方式结合监督（VerbNet标记）和无监督信息，从而获得反映所需信息（动词概念）的适当语义空间。在AffectiveSpace中给定一组动词和大量未标记的实例，将最大化类间分散，并最小化VerbNet实例的类内分散，同时保持所有其他实例的语义相关性。

每个实例表示为$v_i \in A^m$，其是在由随机投影处理之后的m维向量。对于每个动词实例，都有一个标签$y_i \in \{1, \ldots, q\}$，其中q是动词类的数量。然后，类间散布和类内散布矩阵定义如下：

$$S_w = \sum_{j=1}^{q} \sum_{i=1}^{l_j} (v_i - \mu_j)(v_i - \mu_j)^T \tag{5}$$

$$S_b = \sum_{j=1}^{q} l_j (\mu_j - \mu)(\mu_j - \mu)^T \tag{6}$$

其中$\mu_j = \frac{1}{l_j} \sum_{i=1}^{l_j} v_i$（j = 1, 2, ..., q）是j类中样本的均值，l_j是动词的数量

l $i=1$

j类实例和$\mu = \frac{1}{l} \sum_{i=1}^{l} v_i$是所有标记样品的平均值。总散射矩阵AffectiveSpace中的所有实例也被定义：

$$S_t = \sum_{i=1}^{k} (v_i - \mu_k)(v_i - \mu_k)^T \tag{7}$$

其中k是AffectiveSpace中的实例总数，μ_k是所有实例的平均值。我们的目标是找到一个投影矩阵W，将语义空间投射到一个较低维空间，这对动词概念更具辨识力：

$$W^* = \arg\max_{w \in a^{m \times m'}} \frac{W^t S^t W}{|W^T (S_w + \lambda_1 S_t + \lambda_2 I) W|} \tag{8}$$

其中I是单位矩阵，λ_1和λ_2是通过网格搜索获得的控制参数，用于平衡动词判别和语义正则化之间的权衡。最佳解决方案由：

$$(S_w + \lambda_1 S_t + \lambda_2 I) w_{j*} = \lambda_j S_b w_{j*} \qquad j = 1, \ldots, m' \tag{9}$$

其中w_{j*}（j = 1, ..., m'）是对应于（$S_w + \lambda_1 S_t + \lambda_2$的$m'$最大特征值的特征向量）$I)^{-1} S_b$。这里，选择$m' = q - 1$，其中q是总动词原始数。在投影之后，新空间保留了基于语义相关性和动作概念分组
来自AffectNet和VerbNet的信息。

6 评估

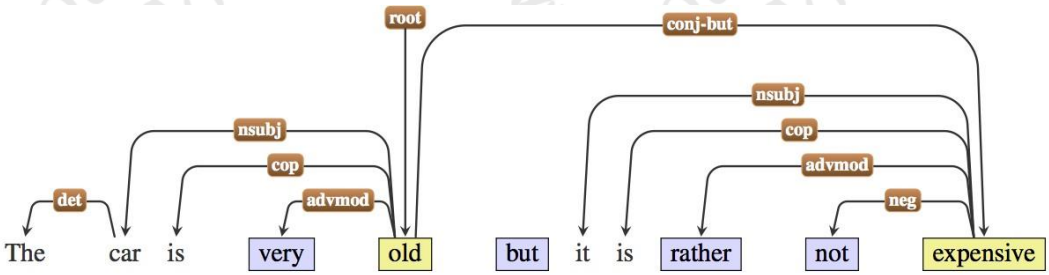
为了对SenticNet 4进行定性评估（既可以作为独立的XML存储库1也可以作为API[2]），我们要求五位注释者判断推断的概念原语的合理性。Cohen的kappa评分为0.84，我们的总体准确率为91%。至于定量评估，我们针对两种众所周知的情绪资源测试了SenticNet 4，即：Blitzer数据集（Blitzer等，2007）和电影评论数据集（Pang and Lee，2005）。

6.1 使用SenticNet执行极性检测

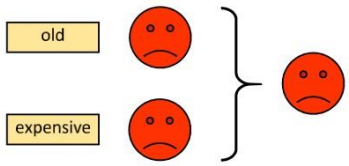
虽然SenticNet可以用作任何其他情感词典，例如概念匹配或概念包模型，但使用知识库进行极性检测任务的正确方法与sentic模式相结合（Poria et al。，2015）。Sentic模式是特定于情感的语言模式，它通过允许情感信息基于条款之间的依赖关系从概念流向概念来推断极性。通过类比电子电路可以最好地说明这种模式背后的主要思想，其中很少“元素”是电荷或信号的“源”，而许多元素通过变换或组合不同信号对信号进行操作。这实现了基本类型的语义处理，其中句子的“含义”仅减少为一个值：其极性。

[1] http://sentic.net/senticnet-4.0.zip

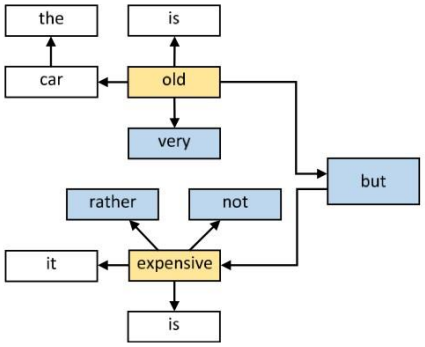
[2] http://sentic.net/api



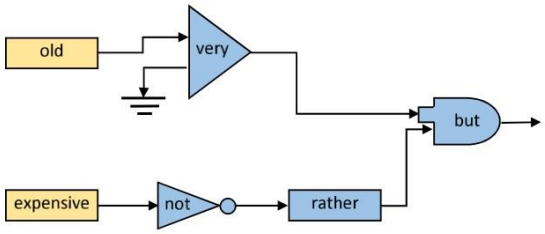
(a) Dependency tree of a sentence.



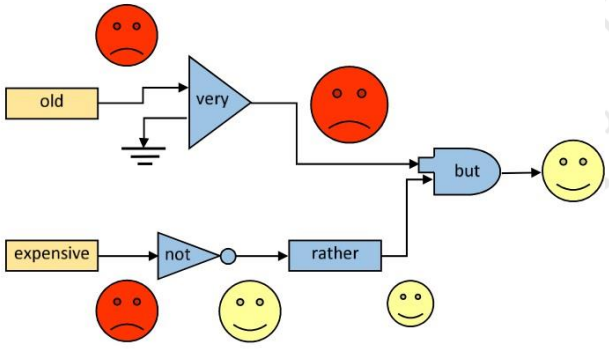
(b) The old way: averaging over a bag of sentiment words. The overall polarity of a sentence is given by the algebraic sum of the polarity values associated with each affect word divided by the total number of words.



(c) The dependency tree of a sentence resembles an electronic circuit: words shown in blue can be thought as a sort of "boolean operations" acting on other words.



(d) The electronic circuit metaphor: sentiment words are "sources" while other words are "elements", e.g., *very* is an amplifier, *not* is a logical complement, *rather* is a resistor, *but* is an OR-like element that gives preference to one of its inputs.



(e) The final sentiment data flow of the "signal" in the "circuit".

图3：在一些模式中，句子的结构类似于电子电路，其中逻辑运算符引导情绪数据流以输出总体极性。

Sentic模式应用于句子的依赖句法树，如图3（a）所示。只有两个具有内在极性的单词以黄色显示;以类似于语境价值变换器的方式修改其他词语含义的词语（Polanyi和Zaenen，2006）以蓝色显示。完全忽略句子结构的基线，以及没有固有极性的单词，如图3（b）所示：剩下的两个单词是负数，因此总极性为负。但是，语法树可以以“电路”的形式重新解释，其中“信号”从一个元素（或子树）流向另一个元素（或子树），如图3（c）所示。在删除不用于极性计算的字（白色）后，获得具有类似电子放大器，逻辑补码和电阻器的元件的电路，如图3（d）所示，

图3（e）说明了工作中的想法：情绪从极性词流过变换器并组合词。在这个例子中，两个带有极性的词是负的。“老”这个词的负面影响被增强器'非常'放大了。然而，“昂贵”这个词的负面影响被否定所反转，所产生的正值被“电阻”减少。最后，两个短语的值通过连接'但'组合，使得整体极性具有与第二组分（正）的相同的符号。

6.2 SenticNet 4与SenticNet 3

Blitzer数据集包含七个不同领域的产品评论。对于每个域，有1,000个正面和1,000个负面评论。在评估SenticNet 4时，我们只使用了电子类别下的评论。从这些中，我们随机抽取了7,210个非中性句子：其中3,800个标记为阳性，3,410个标记为阴性。然后，我们将SenticNet 4的性能与其前身SenticNet 3进行了比较，以便使用sentic模式进行句子级极性检测。结果如表3所示。

表3：Blitzer数据集的比较

骨架	准确性
Sentic Patterns和SenticNet 3	87.0%
Sentic Patterns和SenticNet 4	91.3%

6.3 SenticNet 4与统计方法

电影评论数据集包括从烂番茄[3]收集的1,000个正面和1,000个负面电影评论。最初，Pang和Lee手动将每个评论标记为正面或负面。后来，Socher等人。（Socher et al。，2012；Socher et al。，2013）在句子层面注释了这个数据集。他们从评论中提取了11,855个句子，并使用五种情绪标签的细粒度清单手动标记它们：强阳性，阳性，中性，阴性和强阴性。由于在这项工作中我们只考虑二元分类，我们从数据集中删除了中性句子并合并了锗烷标签。因此，最终数据集包含4,800个正面句子和4,813个正面句子。SenticNet 3和SenticNet 4的分类结果如表4所示。

表4：电影评论数据集的比较

骨架	准确性
Socher等，2012	80.0%
Socher等，2013	85.4%
Sentic Patterns和SenticNet 3	86.2%
Sentic Patterns和SenticNet 4	90.1%

7 结论

从Web上的大量非结构化信息中提取知识是社交媒体营销，品牌定位和财务预测等任务的关键因素。常识推理是情感分析的一个很好的解决方案，但常识知识库的可扩展性是危害概念提取和极性检测效率的主要因素。解决这个问题的第一个可能步骤是根据概念原语概括一些常识知识，这些原语可以捕获大多数自然语言概念的语义变化。

在这项工作中，我们使用了一个层次聚类和降维的集合来自动发现SenticNet中名词和动词概念的原语。这种泛化过程使我们能够在很大程度上扩展常识知识库的覆盖范围，从而提高SenticNet用于句子级极性检测的准确性，与先前版本的资源和最先进的版本相比较统计情绪分析研究。

在未来，我们计划通过基于依赖性的词嵌入以更自动和可扩展的方式发现新的概念原语。特别是，我们将利用内部学习的skip-gram模型的上下文嵌入与标准目标词嵌入相结合，来衡量上下文兼容性以及单词相似性。

[3] http://rottentomatoes.com

参考

Nir Ailon和Bernard Chazelle。2010.减少尺寸的速度更快。ACM的通讯，53（2）：97-104。

Matheus Arau'jo，PollyannaGonc alves，Meeyoung Cha和FabrıcioBenevenuto。2014. iFeel：一种比较和结合情绪分析方法的系统。在WWW，第75-78页。

大卫巴尔杜齐。随机化的共同训练：从皮层神经元到机器学习，再回来。arXiv preprint arXiv：1310.6536。

帕维尔 伯欣。2006.聚类数据挖掘技术的调查。对多维数据进行分组，第25-71页。

Ella Bingham和Heikki Mannila。减少维数的随机投影：图像和文本数据的应用。在ACM SIGKDD，第245-250页。

John Blitzer，Mark Dredze和Fernando Pereira。传记，宝莱坞，繁荣框和混合器：情绪分类的领域适应。在ACL，第7卷，第440-447页。

Felipe Bravo-Marquez，Marcelo Mendoza和Barbara Poblete。2014.用于大型社交数据分析的元级别情绪模型。基于知识的系统，69：86-99。

E*rik Cambria和Amir Hussain。2015. Sentic Computing：基于常识的概念级情感分析框架。Springer，Cham，瑞士。*

Erik Cambria和Bebo White。跳跃NLP曲线：自然语言处理研究综述。*IEEE计算智能杂志，9（2）：48-57。*

Erik Cambria，Daniel Olsher和Dheeraj Rajagopal。SenticNet 3：认知驱动情绪分析的常识和常识知识库。在AAAI，第1515-1521页，魁北克市。

Erik Cambria，Jie Fu，Federica Bisio和Soujanya Poria。2015.AffectiveSpace 2：为概念层面的情绪分析提供情感直觉。在AAAI，第508-514页，奥斯汀。

埃里克坎布里亚。2016.情感计算和情绪分析。IEEE Intelligent Systems，31（2）：102-107。

CıceroNogueirados Santos和MaıraGatti。用于短文本情感分析的深度卷积神经网络。在COLING，第69-78页。

G*illes Fauconnier和Mark Turner。2003.我们思考的方式：概念融合和心灵隐藏的复杂性。基础书籍。*

Gizem Gezici，Rahim Dehkharghani，Berrin Yanikoglu，Dilek Tapucu和Yucel Saygin。2013.Su-sentilab：Twitter中的情绪分析分类系统。在国际语义评估研讨会，第471-477页。

Ray Jackendoff。1976.朝着解释性的语义表示。语言探究，7（1）：89-150。Anil Jain和

Richard Dubes。1988.聚类数据的算法。Prentice-Hall，Inc。

Honglak Lee，Roger Grosse，Rajesh Ranganath和AY Ng。2011年。使用卷积深度信念网络进行无监督学习的等级表示。ACM的通讯，54（10）：95-103。

刘冰。2012.情绪分析和意见挖掘。摩根和Claypool。

Yichao Lu，Paramveer Dhillon，Dean P Foster和Lyle Ungar。2013.通过二次抽样随机化hadamard变换进行更快的岭回归。在神经信息处理系统的进展，第369-377页。

Prem Melville，Wojciech Gryc和Richard D Lawrence。通过将词汇知识与文本分类相结合，对博客的情感分析。在ACM SIGKDD，第1275-1284页。

马文明斯基。1975.代表知识的框架。在Patrick Winston，编辑，计算机视觉心理学。麦格劳希尔，纽约。

Bo Pang和Lillian Lee。看到明星：在评级量表上利用情感分类的阶级关系。在ACL中，第115-124页，Ann Arbor。

Bo Pang和Lillian Lee。2008.意见挖掘和情绪分析。信息检索的基础和趋势，2：1-135。

Bo Pang, Lillian Lee和Shivakumar Vaithyanathan。大拇指？：使用机器学习技术的情感分类。在EMNLP，第10卷，第79-86页。

Livia Polanyi和Annie Zaenen。2006.上下文价格变换器，。“计算文本中的态度和影响：理论与应用”，“信息检索系列”第20卷，第1-10页。施普林格，柏林，德国。

Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio和Amir Hussain。2015.通过动态语言模式进行情感数据流分析。IEEE计算智能杂志，10（4）：26-36。

Soujanya Poria, Erik Cambria和Alexander Gelbukh。2016a。基于深度卷积神经网络的意见挖掘方面提取。基于知识的系统，108：42-49。

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang和Amir Hussain。2016B。融合来自多模式内容的情感分析的音频，视觉和文本线索。Neurocomputing，174：50-59。

Soujanya Poria, Iti Chaturvedi, Erik Cambria和Federica Bisio。2016c。Sentic LDA：对基于方面的情感分析的语义相似性改进LDA。在IJCNN。

Diego Reforgiato Recovery, Valentina Presutti, Sergio Consoli, Aldo Gangemi和Andrea Nuzzolese。2014年
Sentilo：基于框架的情绪分析。认知计算，7（2）：211-225。

David Rumelhart和Andrew Ortony。1977.知识在记忆中的表现。在C Anderson，R Spiro和W Montague，编辑，学校教育和知识的获得。Erlbaum，Hillsdale，NJ。

Tamas Sarlos。通过随机投影改进大矩阵的近似算法。在FOCS中，第143-152页。

Roger Schank。1972年。概念依赖：自然语言理解理论。认知心理学，3：552-631。

卡琳舒勒。2005. VerbNet：广泛覆盖，全面的动词词典。博士论文，宾夕法尼亚大学计算机与信息科学。

编辑巴里史密斯。1988.格式塔理论的基础。慕尼黑和维也纳：Philosophia Verlag。

Richard Socher, Brody Huval, Christopher D Manning和Andrew Y Ng。通过递归矩阵向量空间的语义组合。在EMNLP中，第1201-1211页。

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng和Christopher Potts。2013.针对情感树库的语义组合的递归深度模型。在EMNLP，第1642-1654页。

罗伯特施佩尔和凯瑟琳哈瓦西。2012. ConceptNet 5：关系知识的大型语义网络。在Eduard Hovy，M Johnson和G Hirst，自然语言处理的编辑，理论和应用，第6章Springer。

Carlo Strapparava和Alessandro Valitutti。WordNet-Affect：WordNet的情感扩展。在LREC，第1083-1086页，里斯本。

唐都宇，傅茹，秦兵，刘婷，明周。2014A。Coooolll：用于推特情绪分类的深度学习系统。在SemEval中，第208-212页。

唐都宇，傅茹，杨楠，刘婷，秦兵。2014B。学习针对Twitter情绪分类的特定情绪词嵌入。在ACL中，第1555-1565页。

Joel A Tropp。2011.改进的子采样随机化hadamard变换分析。自适应数据分析的进展，3（01n02）：115-126。

阿莫斯特沃斯基。1977.相似的特征。心理学评论，84（4）：327-352。Anna

Wierzbicka。1996.语义学：Primes和Universals。牛津大学出版社。

Hong Yu和Vasileios Hatzivassiloglou。2003.回答意见问题：将意见与意见分开，并确定意见句子的极性。在EMNLP中，第129-136页。